

Nonparametric imputation by data depth

Pavlo Mozharovskyi*

CREST, Ensai, Université Bretagne Loire

and

Julie Josse

École Polytechnique, CMAP

and

François Husson

IRMAR, Applied Mathematics Unit, Agrocampus Ouest, Rennes

January 12, 2017

Abstract

The presented methodology for single imputation of missing values borrows the idea from data depth — a measure of centrality defined for an arbitrary point of the space with respect to a probability distribution or a data cloud. This consists in iterative maximization of the depth of each observation with missing values, and can be employed with any properly defined statistical depth function. On each single iteration, imputation is narrowed down to optimization of quadratic, linear, or quasiconcave function being solved analytically, by linear programming, or the Nelder-Mead method, respectively. Being able to grasp the underlying data topology, the procedure is distribution free, allows to impute close to the data, preserves prediction possibilities different to local imputation methods (k-nearest neighbors, random forest), and has attractive robustness and asymptotic properties under elliptical symmetry. It is shown that its particular case — when using Mahalanobis depth — has direct connection to well known treatments for multivariate normal model, such as iterated regression or regularized PCA. The methodology is extended to the multiple imputation for data stemming from an elliptically symmetric distribution. Simulation and real data studies positively contrast the procedure with existing popular alternatives. The method has been implemented as an R-package.

Keywords: Elliptical symmetry, Outliers, Tukey depth, Zonoid depth, Nonparametric imputation, Convex optimization.

*The major part of this project has been conducted during the postdoc of Pavlo Mozharovskyi at Agrocampus Ouest (Rennes) granted by *Centre Henri Lebesgue* due to program PIA-ANR-11-LABX-0020-01.

1 Introduction

Following the seminal idea of [Tukey \(1974\)](#), the concept of data depth has developed to a powerful statistical methodology allowing description of data w.r.t. location, scale, and shape based on a multivariate ordering. Today, it finds numerous applications in multivariate data analysis ([Liu et al., 1999](#)), statistical quality control ([Liu and Singh, 1993](#)), classification ([Jörnsten, 2004](#), [Lange et al., 2014](#)), multivariate risk measurement ([Cascos and Molchanov, 2007](#)), robust linear programming ([Bazovkin and Mosler, 2015](#)), *etc.* Being able to exploit topological properties of the data in a nonparametric way, statistical depth function further proves to be suited for imputing missing values while being connected to the state of the art methods. We start with a brief background on imputation, followed by the proposal.

1.1 Background on missing values

The problem of missing values exists since the earliest attempts of exploiting data as a source of knowledge as it lies intrinsically in the process of obtaining, recording, and preparation of the data itself. To exploit all the information present in the data set, a statistical method may be adapted to missing values, but this requires developing such a one for each estimator and inference of interest. A more universal way is to impute missing data first, and then apply the statistical method of interest to the completed data set ([Little and Rubin, 2002](#)). Lastly, the multiple imputation has gained a lot of attention: for a data set containing missing values a number of completed data sets is generated reflecting uncertainty of the imputation process, which enables not only estimating the parameter of interest but also drawing an inference on it ([Rubin, 1996](#)).

Development of many statistical methods started with the natural normality assumption, and imputation is not an exception here. For a multivariate normal distribution, single imputation of the observations containing missing values can be performed by imputing the missing values with conditional mean, where conditioning is w.r.t. observed values. This procedure makes use of the expectation-maximization algorithm ([Dempster et al., 1977](#)) to estimate mean and covariance matrix; see also [Little and Rubin \(2002\)](#). By that, imputation inherits sensitivity to outliers and to (near) low-rank covariance matrix. To deal with these problems, uncountable extensions of the EM framework have been developed. Another way is to directly assume low-rank of the underlying covariance, which is followed by the methods based on the principal component analysis (PCA), closely connected to the matrix completion literature, see [Josse and Husson \(2012\)](#), [Hastie et al. \(2015\)](#). These methods aim at denoising the data and suppose its special structure, knowledge (or estimation) of the rank of data, of the shape of outliers, and noise to be normal. Both groups of methods impute on low-dimensional affine subspaces and — by that — are sensitive to even slight deviations from normality (or ellipticity in general) and ignore the geometry of data. Extension of single imputation methods to more general densities consists in such nonparametric techniques as k -nearest neighbors (k NN) (see [Troyanskaya et al., 2001](#), and references there in) and random forest ([Stekhoven and Bühlmann, 2012](#)) imputation. When properly tuned, these methods capture the locality of the data, but fail to extrapolate and thus can be inappropriate under the missing at random mechanism (MAR, [Seaman et al., 2013](#)), *i.e.* exhibit in some sense superfluously local behavior.

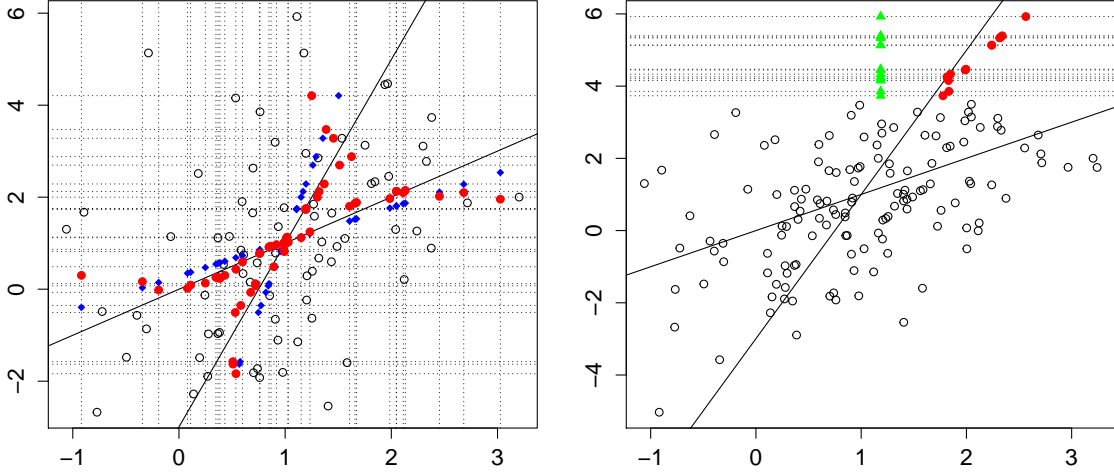


Figure 1: Bivariate normal distribution with 30% MCAR (left) and with MAR in second coordinate for values > 3.5 (right); imputation using maximum zonoid depth (red), conditional mean imputation using EM estimates (blue), and random forest imputation (green).

1.2 Proposal

Current article fills the existing gap on the level of single imputation for general elliptically symmetric distribution families. We suggest a nonparametric framework that is able to grasp data topology and, to some degree, even reflect finite sample deviations from ellipticity. For the purpose of single imputation of missing values, we resort to the idea of statistical depth function $D(\mathbf{x}|X)$ — a measure of centrality of an arbitrary point $\mathbf{x} \in \mathbb{R}^d$ w.r.t. a d -variate random vector X . Given such a measure, we propose to maximize it for a point having missing values conditioned on its observed values. For each point containing missing values, the procedure is repeated iteratively to achieve stability of the solution. In this framework, the properties of the imputation are to great extent defined by the chosen notion of data depth function (see Section 2). A nonparametric data depth allows for imputation close to the data geometry, still accounting for their global features due to the center anchoring. This makes a distinction w.r.t. using fully local imputation methods like k NN or random forest imputation. In addition, data depth allows for robust imputation both in sense of outliers (i.e. disregarding outliers when imputing points closer to center) and distribution (i.e. not masking outliers). As not using it, depth-based imputation avoids problems connection with estimation of the covariance matrix.

We employ three depth functions: Mahalanobis depth, which is a natural extension of the Mahalanobis distance and is here because of its connections to standard imputation frameworks (see Section 4); zonoid depth, which imputes by the average of the maximal number of equally weighted observations; Tukey depth, which maximizes the infimum of the portion of points contained in the closed halfspace including the imputed point. For a well defined imputation, these depth notions require two, one, and zero first moments of the underlying probability measure, respectively. We highlight most important properties of the depth-based imputation right below by means of examples.

Regard Figure 1 (left), where 150 points stemming from a bivariate normal distribution having mean $\boldsymbol{\mu}_1 = (1, 1)'$ and covariance $\boldsymbol{\Sigma}_1 = ((1, 1)', (1, 4)')$ are plotted, with 30% of the entries having been removed in both variables due to the missing completely at random (MCAR) mecha-

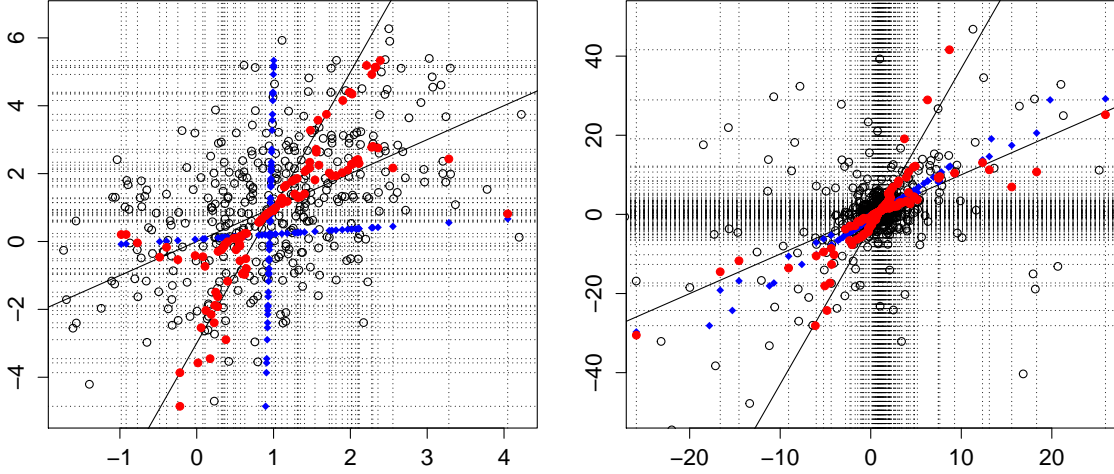


Figure 2: Mixture of bivariate normal (425 points, 15% MCAR) and Cauchy (75 point) samples (left) and 1000 bivariate Cauchy distributed points with 15% MCAR (right). Imputation using maximum Tukey depth (red) and conditional mean imputation using EM estimates (blue).

nism; points with one missing entry are denoted by dotted lines. EM-based imputation is depicted in blue. The imputed points are conditional means and lie exactly on the regression lines estimate; their distribution versions are plotted as the two lines. Zonoid depth-based imputation, pictured in red, reflects the idea that the sample may not necessarily be normal, with this unsureness more stressed on the fringe of the data cloud, where imputed points deviate from conditional mean towards the unconditional one. Indeed, away from the data center, imputation by data depth is closer to nonparametric imputation, which accounts for the local nature of data. On the other hand, different to these methods, depth imputation preserves prediction ability. In Figure 1 (right) for the same sample first coordinate has been removed for (14) points having value > 3.5 in the second coordinate (MAR mechanism). The depth-based imputation manages to extrapolate to predict missing values, while random forest imputation performs as expected rather poorly.

Being robust both in sense of outliers and heavy tails concerns the two following cases: First, if data is polluted by outliers but coordinates are missing for a representative point, the outliers should not alter the imputation. In Figure 2, left we plot — zoomed in — 500 points. 425 of them stem from the same normal distribution as above, with 15% of entries MCAR; another 75 are outliers drawn from the Cauchy distribution with the same center and shape matrix and having no missing values. As expected, conditional mean based on EM estimates (depicted in blue) imputes rather randomly. Depth-based imputation using Tukey depth imputes in a robust way, close to (distribution) regression lines, reflecting geometry of data. Second, if missing values belong to the polluting distribution (or if the entire data generating process is heavy-tailed), the imputation should reflect this heavy-tailness not to mask a possible outlier. For 1000 points from Cauchy distribution having 15% of values MCAR, imputation by Tukey depth and the EM-based one (for comparison) are shown in Figure 2, right. Depth-based respects the general ellipticity and imputes close to distribution regression lines.

The rest of the paper is organized as follows. In Section 2 we state some important definitions concerning data depth and establish the notation. Section 3 enlightens the proposed methodology of the depth-based imputation, regards its theoretical properties, and suggests optimization techniques. Section 4 is devoted to the theoretical investigation of the special case when employing

Mahalanobis depth bridging the proposed imputation technique with regression and PCA imputation. Section 5 provides a comparative simulation and real data study. Section 6 extends the proposed approach to multiple imputation. Section 7 gathers some useful remarks.

2 Notation and background on data depth

For a data set in \mathbb{R}^d ($n \times d$ matrix) $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} = (\mathbf{X}_{obs}, \mathbf{X}_{miss})$, we denote by \mathbf{X}_{obs} and \mathbf{X}_{miss} its observed and missing part, respectively, *i.e.* there exists an indicator $n \times d$ matrix \mathbf{M} such that $\mathbf{X} = \mathbf{X}_{obs} \cdot (\mathbf{1}_{n \times d} - \mathbf{M}) + \mathbf{X}_{miss} \cdot \mathbf{M}$ in the sense of elementwise multiplication. For a point $\mathbf{x} \in \mathbb{R}^d$, we denote $miss(\mathbf{x})$ and $obs(\mathbf{x})$ the sets of its coordinates containing missing, respectively observed values, denoting $|miss(\mathbf{x})|$ and $|obs(\mathbf{x})|$ the corresponding cardinalities, restricting to $|miss(\mathbf{x})| + |obs(\mathbf{x})| = d$. Following the same logic, we write $miss(i)$ and $obs(i)$ for a point $\mathbf{x}_i \in \mathbf{X}$ or even just $miss$ and obs if no confusion arises.

Under missing completely at random (MCAR) we understand the mechanism where each value of \mathbf{X} has the same missing probability.

Being in the core of the theoretical results derived in Section 3 elliptical distribution is defined as follows (see Fang et al. (1990), and Liu and Singh (1993) in the data depth context).

Definition 1. Random vector $X \in \mathbb{R}^d$ is distributed elliptically symmetric with strictly decreasing density or equivalently as $\mathcal{E}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, F)$ if it is distributed as $X \stackrel{D}{=} \boldsymbol{\mu}_X + R\boldsymbol{\Lambda}U$ with $R \in \mathbb{R}_+$ being a nonnegative random variable stemming from F possessing strictly decreasing density, U uniformly distributed on S^{d-1} , $\boldsymbol{\mu}_X \in \mathbb{R}^d$, and $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}'$.

Below we briefly state definitions covering necessary material on data depth. Following the pioneering idea of Tukey (1974), statistical data depth function is a mapping

$$\mathbb{R}^d \times \mathcal{M} \rightarrow [0, 1] : (\mathbf{x}, P) \mapsto D(\mathbf{x}|P),$$

with \mathcal{M} being a subset of \mathcal{M}_0 , the set of all probability measures on $(\mathbb{R}^d, \mathcal{B})$. It measures closeness of \mathbf{x} to the center of P . Further, for a given point \mathbf{x} and a random vector X coming from the probability distribution $P \in \mathcal{M}$ we denote the depth $D(\mathbf{x}|X)$, and $D(\mathbf{x}|\mathbf{X})$ its empirical version. Zuo and Serfling (2000) developed axiomatic for the depth, which requires a proper depth function to be affine invariant, maximal at the P 's center of symmetry, non-increasing on a ray starting from any deepest point, and vanish in infinity. Additionally one may require quasiconcavity, a restriction which will prove to be useful computationally below. Upper-level set of the depth function to level α is called a depth (α -trimmed) region $D_\alpha(X) = \{\mathbf{x} \in \mathbb{R}^d : D(\mathbf{x}|X) \geq \alpha\}$. For $\alpha \in [0, 1]$, depth regions form a family of nested set-valued statistics, which describes X w.r.t. location, scatter, and shape.

A number of depth notions emerged during the last decades; below we give definitions of those three being used for mean of imputation in the present article.

Definition 2. Mahalanobis (1936) depth of $\mathbf{x} \in \mathbb{R}^d$ w.r.t. X is defined as

$$D^M(\mathbf{x}|X) = (1 + (\mathbf{x} - \boldsymbol{\mu}_X)' \boldsymbol{\Sigma}_X^{-1} (\mathbf{x} - \boldsymbol{\mu}_X))^{-1},$$

where $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ are any location and shape estimates of X .

In the empirical version, μ_X and Σ_X are replaced by appropriate estimates of location and shape, which throughout the article are taken as the moment estimates $\mu_X = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ and $\Sigma_X = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \mu_X)(\mathbf{x}_i - \mu_X)'$, respectively.

[Koshevoy and Mosler \(1997\)](#) define a zonoid trimmed region — for $\alpha \in (0, 1]$ — as

$$D_\alpha^z(X) = \left\{ \int_{\mathbb{R}^d} \mathbf{x} g(\mathbf{x}) dP(\mathbf{x}) : g : \mathbb{R}^d \mapsto \left[0, \frac{1}{\alpha}\right] \text{ measurable and } \int_{\mathbb{R}^d} g(\mathbf{x}) dP(\mathbf{x}) = 1 \right\}$$

and for $\alpha = 0$ as

$$D_0^z(X) = \text{conv}(\text{supp}(X)) .$$

where $\text{supp}(X)$ denotes the support of X and $\text{conv}(A)$ denotes the smallest convex set containing A . Its empirical version can be defined as

$$D_\alpha^{z(n)}(\mathbf{X}) = \left\{ \sum_{i=1}^n \lambda_i \mathbf{x}_i : \sum_{i=1}^n \lambda_i = 1, \lambda_i \geq 0, \alpha \lambda_i \leq \frac{1}{n} \forall i \in \{1, \dots, n\} \right\} .$$

Definition 3. *Zonoid depth of \mathbf{x} w.r.t. X is defined as*

$$D^z(\mathbf{x}|X) = \begin{cases} \sup\{\alpha : \mathbf{x} \in D_\alpha^z(X)\} & \text{if } \mathbf{x} \in \text{conv}(\text{supp}(X)), \\ 0 & \text{otherwise.} \end{cases}$$

For a comprehensive reference on the zonoid depth the reader is referred to [Mosler \(2002\)](#).

Zonoid depth tends to represent \mathbf{x} as the average of maximum number of equally weighted points, *i.e.* as a weighted mean, which opens a variety of connected methods including the entire class of the weighted mean depths, see [Dyckerhoff and Mosler \(2011\)](#).

Definition 4. *Tukey (1974) depth of \mathbf{x} w.r.t. X is defined as*

$$D^T(\mathbf{x}|X) = \inf\{P(H) : H \text{ a closed halfspace, } \mathbf{x} \in H\} .$$

In empirical version, probability is substituted by the portion of \mathbf{X} . Exploiting solely the data geometry and optimizing over the indicator loss, Tukey depth is fully nonparametric, highly robust, and does not require moment assumptions on X . For more information on Tukey depth and the corresponding trimmed regions see [Donoho and Gasko \(1992\)](#) and [Hallin et al. \(2010\)](#).

3 Imputation by depth maximization

3.1 Main idea

Let's start by regarding one of the simplest methods to impute missing values making use of the following iterative regression imputation scheme: (1) initialize missing values arbitrary, using mean imputation for instance; (2) impute missing values in one variable by the values predicted by the regression model of this variable with the resting variables taken as explanatory ones, (3) iterate through variables containing missing values till convergence.

Most common imputation methods are based on similar approaches: [Templ et al. \(2011\)](#) use the same (robustified) iterative method, and [Josse and Husson \(2012\)](#) and [Hastie et al. \(2015\)](#) use iterative (thresholded) singular value decomposition to impute the missing entries, whereas [Stekhoven and Bühlmann \(2012\)](#) iteratively replace missing cells with values fitted by random

forest. This scheme can even be traced back to the earliest solution to deal with missing values suggested by [Healy and Westmacott \(1956\)](#) and is at the root of EM based algorithms (implemented *e.g.* by means of the SWEEP operator for normal data) to estimate parameters despite missing values ([Little and Rubin, 2002](#)).

In the described above procedure, on each step, each point \mathbf{x}_i having missing values on a coordinate j is imputed with $\mathbf{y}_{i,j}$ the univariate conditional mean $\mathbb{E}[X|X_{\{1,\dots,d\}\setminus\{j\}} = \mathbf{x}_{i,\{1,\dots,d\}\setminus\{j\}}, \boldsymbol{\mu}_X = \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_X]$. After convergence, each point \mathbf{x}_i with missing values on $\text{miss}(i)$ is imputed with the multivariate conditional mean

$$\begin{aligned} & \mathbb{E}[X|X_{\text{obs}(i)} = \mathbf{x}_{i,\text{obs}(i)}, \boldsymbol{\mu}_X = \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_X] \\ &= \boldsymbol{\mu}_{X_{\text{miss}(i)}} + \boldsymbol{\Sigma}_{X_{\text{miss}(i),\text{obs}(i)}} \boldsymbol{\Sigma}_{X_{\text{obs}(i),\text{obs}(i)}}^{-1} (\mathbf{x}_{i,\text{obs}(i)} - \boldsymbol{\mu}_{X_{\text{obs}(i)}}). \end{aligned} \quad (1)$$

The last expression is the closed-form solution to

$$\min_{\mathbf{z}_{\text{miss}(i)} \in \mathbb{R}^{|\text{miss}(i)|}, \mathbf{z}_{\text{obs}(i)} = \mathbf{x}_{\text{obs}(i)}} d^{Mah}(\mathbf{z}, \boldsymbol{\mu}_X | \boldsymbol{\Sigma}_X)$$

with $d^{Mah}(\mathbf{z}, \boldsymbol{\mu}_X | \boldsymbol{\Sigma}_X) = (\mathbf{z} - \boldsymbol{\mu}_X)' \boldsymbol{\Sigma}_X^{-1} (\mathbf{z} - \boldsymbol{\mu}_X)$ being the squared Mahalanobis distance from \mathbf{z} to $\boldsymbol{\mu}_X$. Minimizing Mahalanobis distance can be seen as maximizing a centrality measure — the Mahalanobis depth

$$\max_{\mathbf{z}_{\text{miss}(i)} \in \mathbb{R}^{|\text{miss}(i)|}, \mathbf{z}_{\text{obs}(i)} = \mathbf{x}_{\text{obs}(i)}} D^{Mah}(\mathbf{z} | \mathbf{X}).$$

We generalize this principle to the notion of statistical depth function.

We propose a unified framework based on the statistical data depth function. Having a sample \mathbf{X} , impute a point \mathbf{x} containing missing coordinates with the point \mathbf{y} maximizing data depth conditioned on observed values \mathbf{x}_{obs} . This direct extension of the idea of conditional mean imputation to data depth can be expressed as

$$\mathbf{y} = \underset{\mathbf{z}_{\text{miss}} \in \mathbb{R}^{|\text{miss}|}, \mathbf{z}_{\text{obs}} = \mathbf{x}_{\text{obs}}}{\text{argmax}} D(\mathbf{z} | \mathbf{X}). \quad (2)$$

The use of equation (2) is limited to strictly quasiconcave continuous and nowhere vanishing depth notions, which is not the case for intrinsically nonparametric depths best reflecting the data geometry. A solution to (2) can trivially be nonunique, as fitting to the finite-sample data topology the depth can be non-continuous. In addition, the value of the depth function may become zero immediately beyond the convex hull of the support of the distribution, which is just $\text{conv}(\mathbf{X})$ for a finite sample. To circumvent these problems, we suggest to impute \mathbf{x} having missing values with \mathbf{y} :

$$\mathbf{y} = \text{ave} \left(\underset{\mathbf{u} \in \mathbb{R}^d, \mathbf{u}_{\text{obs}} = \mathbf{x}_{\text{obs}}}{\text{argmin}} \{ \|\mathbf{u} - \mathbf{v}\| \mid \mathbf{v} \in D_{\alpha^*}(\mathbf{X}) \} \right), \quad (3)$$

where

$$\alpha^* = \inf_{\alpha \in (0;1)} \{ \alpha \mid D_{\alpha}(\mathbf{X}) \cap \{ \mathbf{z} \mid \mathbf{z} \in \mathbb{R}^d, \mathbf{z}_{\text{obs}} = \mathbf{x}_{\text{obs}} \} = \emptyset \}. \quad (4)$$

Given a sample containing missing data $\mathbf{X} = (\mathbf{X}_{\text{obs}}, \mathbf{X}_{\text{miss}})$. Start with an arbitrary initialization of all missing values, by imputing with unconditional mean say. Then, for each observation containing missing coordinates, impute them according to (3), and iterate. This imputation by iterative maximization of depth can be summarized as Algorithm 1.

Algorithm 1 Single imputation

```
1: function IMPUTE.DEPTH.SINGLE( $\mathbf{X}$ )
2:    $\mathbf{Y} \leftarrow \mathbf{X}$ 
3:    $\boldsymbol{\mu} \leftarrow \hat{\boldsymbol{\mu}}^{(obs)}(\mathbf{X})$  ▷ Calculate mean ignoring missing values
4:   for  $i = 1 : n$  do
5:     if  $miss(i) \neq \emptyset$  then
6:        $\mathbf{y}_{i,miss(i)} \leftarrow \boldsymbol{\mu}_{miss(i)}$  ▷ Impute with unconditional mean
7:    $I \leftarrow 0$ 
8:   repeat ▷ Iterate until convergence or maximal iteration
9:      $I \leftarrow I + 1$ 
10:     $\mathbf{Z} \leftarrow \mathbf{Y}$ 
11:    for  $i = 1 : n$  do
12:      if  $miss(i) \neq \emptyset$  then ▷ Impute with maximum depth
13:         $\alpha^* \leftarrow \inf_{\alpha \in (0;1)} \{ \alpha \mid D_\alpha(\mathbf{X}) \cap \{ \mathbf{z} \mid \mathbf{z} \in \mathbb{R}^d, \mathbf{z}_{obs} = \mathbf{y}_{i,obs(i)} \} = \emptyset \}$ 
14:         $\mathbf{y}_i \leftarrow \text{ave}(\arg \min_{\mathbf{u} \in \mathbb{R}^d, \mathbf{u}_{obs} = \mathbf{y}_{i,obs(i)}} \{ \|\mathbf{u} - \mathbf{v}\| \mid \mathbf{v} \in D_{\alpha^*}(\mathbf{X}) \})$ 
15:    until  $\max_{i \in \{1, \dots, n\}, j \in \{1, \dots, d\}} |\mathbf{y}_{i,j} - \mathbf{z}_{i,j}| < \epsilon$  or  $I = I_{max}$ 
16:    return  $\mathbf{Y}$ 
```

After the stopping criterion has been reached, one can expect that for each point initially containing missing values, it holds $\mathbf{x}_i = \arg\max_{\mathbf{z}_{obs} = \mathbf{x}_{obs}} \min_{\mathbf{u} \in S^{d-1}} |\{k : \mathbf{x}'_k \mathbf{u} \geq \mathbf{z}'_k \mathbf{u}, k \in \{1, \dots, n\}\}|$ when employing the Tukey depth. So the imputation is performed due to the maximin principle based on criteria involving indicator functions, which implies robustness of the solution. When using zonoid depth, each such \mathbf{x}_i is imputed by the average of the maximum number of possibly most equally weighted points. W.r.t. \mathbf{X} , this is a weighted mean imputation, which has connection to the methods of local nature such as the k NN imputation, and allows for gaining additional insights into data geometry by inspecting the optimal weights, constituting the Lagrange multipliers; see Section 3.3 for the detailed discussion. With Mahalanobis depth, each \mathbf{x}_i with missingness is imputed by the conditional mean (1) and thus lies in the single-output regression hyperplane $\mathbf{X}_{\cdot,j}$ on $\mathbf{X}_{\cdot, \{1, \dots, d\} \setminus \{j\}}$ for all $j \in miss(i)$ and in general in the $(d - |miss(i)|)$ -dimensional multiple-output regression subspace $\mathbf{X}_{\cdot, miss(i)}$ on $\mathbf{X}_{\cdot, obs(i)}$; such a subspace is obtained as the intersection of the single-output regression hyperplanes corresponding to missing coordinates. Among others, this yields further properties and connections to covariance determinant and the regularized PCA imputation by Josse and Husson (2012), which is regarded in detail in Section 4.

3.2 Properties

Though being of different nature, suggested depth-based framework maintains some of the desirable properties of the EM-based imputation, since the imputation converges to center of the conditional distribution in the settings specified in Theorems 1 and 2. Different to EM, we avoid the second and first moment assumptions due to the use zonoid and Tukey depths. Together with Mahalanobis depth, the choice of depths becomes canonical in the sense that finiteness of first two (Mahalanobis depth), one (zonoid depth), and no (Tukey depth) moments is required.

Theorem 1 shows that for any elliptical distribution, imputation of one point only converges to the center of the conditional distribution when conditioning on the observed values.

Theorem 1 (One row consistency). *Let $\mathbf{X}^{(n)} = (\mathbf{X}_{obs}^{(n)}, \mathbf{x})$ be a sequence of data sets in \mathbb{R}^d with $d \geq 2$ consisting of a point $\mathbf{x} = (\mathbf{x}_{obs}, \mathbf{x}_{miss})$ and $\mathbf{X}_{obs}^{(n)}$ being n points sampled from an $\mathcal{E}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X), f$. Then for $n \rightarrow \infty$ for Tukey, zonoid (\mathcal{E} possesses finite 1st moment), and Mahalanobis (\mathcal{E} possesses finite 2nd moment) depths*

$$\mathbf{y}_{miss} = \boldsymbol{\mu}_{X_{miss}} + \boldsymbol{\Sigma}_{X_{miss}, obs} \boldsymbol{\Sigma}_{X_{obs}, obs}^{-1} (\mathbf{x}_{obs} - \boldsymbol{\mu}_{X_{obs}})$$

is a stationary point of Algorithm 1.

Theorem 1 is illustrated in Figure 3 for a bivariate sample stemming from the Cauchy distribution, where a point is imputed using Tukey depth. The kernel density estimate of the imputed values over 10 000 repetitions resembles the Gaussian curve and approaches the target value being the center of the population conditional distribution.

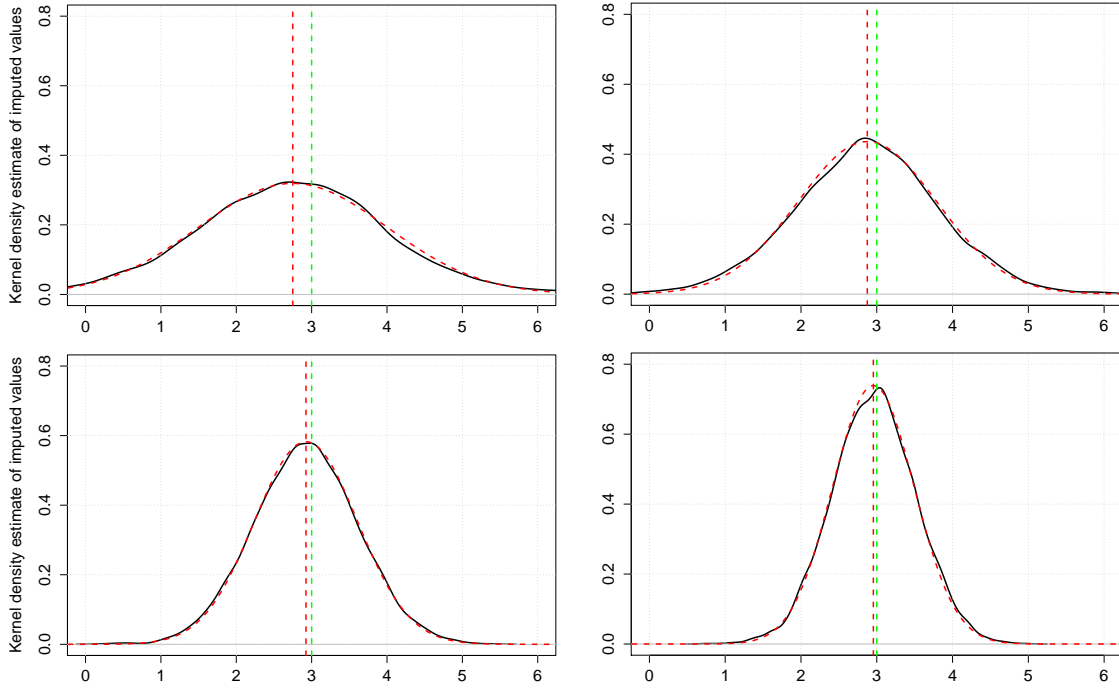


Figure 3: Kernel density estimate (solid) and the best approximating Gaussian curve (dashed) over 10 000 repetitions of the imputation of a single point having one missing coordinate. Sample of size 100 (top, left), 200 (top, right), 500 (bottom, left), 1000 (bottom, right) is drawn from the Cauchy distribution with location and scatter parameters $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ from the introduction. The population conditional center given the observed value equals 3.

Theorem 2 states that if missing values constitute a portion of the sample but concentrate in a single variable, the imputed values converge to the center of the conditional distribution when conditioning on the observed values.

Theorem 2 (One column consistency). *Let $\mathbf{X}^{(n)} = (\mathbf{X}_{obs}^{(n)}, \mathbf{X}_{miss}^{(n)})$ be a sequence of data sets with $\mathbf{x}_{miss}^{(n)}$ only in coordinate j following MCAR mechanism, sampled from an $\mathcal{E}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, F)$. Then for $n \rightarrow \infty$ for Tukey, zonoid (\mathcal{E} possesses finite 1st moment), and Mahalanobis (\mathcal{E} possesses finite 2nd moment) depths for all points having missing values*

$$\mathbf{y}_j = \boldsymbol{\mu}_{X_j} + \boldsymbol{\Sigma}_{X_j, obs} \boldsymbol{\Sigma}_{X_{obs, obs}}^{-1} (\mathbf{x}_{obs} - \boldsymbol{\mu}_{X_{obs}})$$

with $obs = \{1, \dots, d\} \setminus \{j\}$ is a stationary point of Algorithm 1.

The statements of both theorems are valid for $n \rightarrow \infty$. Nevertheless, for any $\epsilon > 0$ there exists N such that for all $n > N$ ϵ -neighborhood of the corresponding limit from Theorems 1 and 2 is a stationary point of Algorithm 1.

3.3 Optimization in a single iteration

Due to Algorithm 1, the only imputation step repeated iteratively consists in imputing point \mathbf{x} having missing coordinates $miss(\mathbf{x})$ with \mathbf{y} by maximizing its depth $D(\mathbf{z}|\mathbf{X})$ w.r.t. the completed data set \mathbf{X} conditioned on \mathbf{x}_{obs} . Note that the function $f(\mathbf{z}_{miss}(\mathbf{x})) := D(\mathbf{z}|\mathbf{X}, \mathbf{z}_{obs}(\mathbf{x}) = \mathbf{x}_{obs})$ is quadratic for Mahalanobis depth, continuous inside $\text{conv}(\mathbf{X})$ for zonoid depth, and step-wise discrete there for Tukey depth, with the single solution to a convex maximization problem in all the three cases. For a trivariate Gaussian sample, $f(\mathbf{z}_{miss}(\mathbf{x}))$ is depicted in Figure 4

For Mahalanobis depth, equation (2) has closed form: $\mathbf{y}_{miss} = \boldsymbol{\mu}_{X_{miss}} + \boldsymbol{\Sigma}_{X_{miss, obs}} \boldsymbol{\Sigma}_{X_{obs, obs}}^{-1} (\mathbf{x}_{obs} - \boldsymbol{\mu}_{X_{obs}})$. In case $\boldsymbol{\Sigma}_X$ is singular, one can work in the linear subspace of eigenvectors with positive eigenvalues. Defined this way Mahalanobis depth is sensitive to outliers, which can be compensated for by estimating $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$ in a robust way, using minimum covariance determinant (MCD, see Rousseeuw and Van Driessen, 1999) say.

Zonoid depth is continuous inside the convex hull of the sample, and thus optimization w.r.t. (2) can be used directly. By construction zonoid depth can be represented as a problem of linear programming, see Mosler (2002) for details. To account for the missingness, a modification is necessary, which consists in removing constraints corresponding to the point's missing coordinates:

$$\begin{aligned} \min \gamma \quad \text{s. t.} \quad & \mathbf{X}_{obs(\mathbf{x})} \boldsymbol{\lambda} = \mathbf{x}_{obs}, \\ & \boldsymbol{\lambda} \mathbf{1}_n = 1, \\ & \gamma \mathbf{1}_n - \boldsymbol{\lambda} \geq \mathbf{0}_n, \\ & \boldsymbol{\lambda} \geq \mathbf{0}_n. \end{aligned} \tag{5}$$

Here $\mathbf{X}_{obs(\mathbf{x})}$ the completed $n \times |obs(\mathbf{x})|$ data matrix containing columns corresponding only to nonmissing coordinates of \mathbf{x} and $\mathbf{1}_n$ (respectively $\mathbf{0}_n$) a vector of ones (respectively zeros) of length n . Imputation is finally performed with the $\boldsymbol{\lambda}$ -weighted average

$$\mathbf{y}_{miss} = \mathbf{X}'_{miss(\mathbf{x})} \boldsymbol{\lambda}.$$

Investigating $\boldsymbol{\lambda} = \{\lambda_1, \dots, \lambda_n\}$, the weights applied to the imputed point, may give additional insights concerning its positioning w.r.t. the sample. Thus, from (5) it is clear that in case it is solvable, the number of nonnegative weights $|\{i : \lambda_i > 0, i \in \{1, \dots, n\}\}| = m + 1$, $1 \leq m + 1 \leq n$, where usually m of them are equal to the solution γ^* , and one is ≥ 0 and $\leq \gamma^*$. This means that \mathbf{y}_{miss} can be seen as the average of m points of the sample very slightly shifted

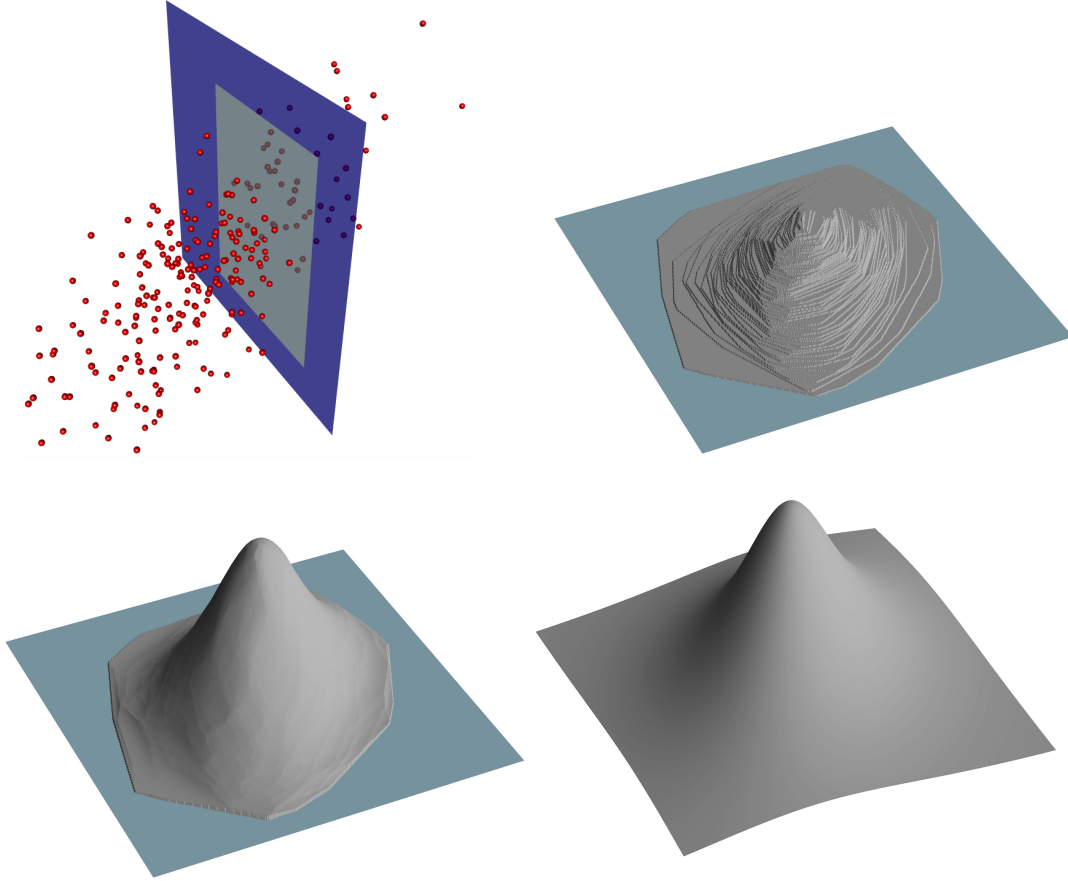


Figure 4: A Gaussian sample consisting of 250 points and a hyperplane of two missing coordinates (top, left), and the function $f(z_{miss(x)})$ to be optimized on each single iteration of Algorithm 1, for the identified (smaller) rectangle, for Tukey (top, right), zonoid (bottom, left), and Mahalanobis (bottom, right) depth.

by an $(m + 1)$ st one. These $m + 1$ points constitute the part of the imputed sample “responsible” for imputation of x_{miss} , and can be readily identified by strictly positive weights λ_i .

Due to the combinatorial nature of the discrete Tukey depth function another approach is needed. We employ the Nelder-Mead downhill-simplex method, which is called $2d$ times with differing starting points; the results are averaged after. Although this slightly deviates from (3), it works stably in practice and one can expect convergence to (3) for a continuous density. For the purity of experiment, in the study of Section 5 we always compute Tukey depth exactly following Dyckerhoff and Mozharovskiy (2016), although to avoid computational burden approximation through random directions (Dyckerhoff, 2004) is recommended.

In the above paragraphs, we were considering the problem of solving (3) inside $\text{conv}(\mathbf{X})$. Zonoid and Tukey depths equal $1/n$ on the $\text{conv}(\mathbf{X})$ and 0 everywhere beyond this. While (3) deals with this situation, for a finite sample this means that points with missingness having maximal value in at least one of the existing coordinates will never move from the initial imputation because they will become vertices of the $\text{conv}(\mathbf{X})$. For the similar reason, other points to be imputed and lying exactly on the $\text{conv}(\mathbf{X})$ will have suppressed dynamics. As such points are

not numerous and thus need quite a substantial deviation to influence imputation quality, we impute them — during a few first iterations — using the spatial depth function (Vardi and Zhang, 2000) that is everywhere nonnegative. This resembles the idea the so-called “outsider treatment” introduced by Lange et al. (2014). To make spatial depth affine invariant, covariance matrix is used, which is taken as the moment estimator for zonoid depth and as the MCD estimator (with parameter 0.5) for Tukey depth.

4 Special case: Mahalanobis depth

As announced in Section 3, the case of Mahalanobis depth imputation is additionally interesting, also due to its relation to existing methods. In this section we consider two points: its connection to the minimization of the covariance determinant, and correspondence of the final imputation to the iterative regression and the regularized PCA imputation. Proposition 1 gives the first insight into the change of the determinant of the sample covariance matrix by stating that it decreases when imputing a single point.

Proposition 1 (Imputation by conditional mean reduces covariance determinant). *Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a sample in \mathbb{R}^d with $\boldsymbol{\mu}_{\mathbf{X}} = 0$ and invertible $\boldsymbol{\Sigma}_{\mathbf{X}}$. Further, for some $k \in \{1, \dots, n\}$, let $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ with $\mathbf{y}_i = \mathbf{x}_i$ for all $i \in \{1, \dots, n\} \setminus \{k\}$, $\mathbf{y}_{k, \text{miss}(k)} = \boldsymbol{\Sigma}_{\text{miss}(k), \text{obs}(k)} \boldsymbol{\Sigma}_{\text{obs}(k), \text{obs}(k)}^{-1} \mathbf{x}_{k, \text{obs}(k)}$, and $\mathbf{y}_{k, \text{obs}(k)} = \mathbf{x}_{k, \text{obs}(k)}$, such that $\mathbf{y}_k \neq \mathbf{x}_k$. Then $|\boldsymbol{\Sigma}_{\mathbf{Y}}| < |\boldsymbol{\Sigma}_{\mathbf{X}}|$.*

Lemma 1 points out that the relationship in Proposition 1 is quadratic, and this fact is later used to prove point (2) of Theorem 3.

Lemma 1. *Let $\mathbf{X}(\mathbf{y}) = (\mathbf{x}_1, \dots, (\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,|\text{obs}(i)|}, \mathbf{y}')', \dots, \mathbf{x}_n)'$ be a $n \times d$ matrix with $\boldsymbol{\Sigma}_{\mathbf{X}}(\mathbf{y})$ invertible for all $\mathbf{y} \in \mathbb{R}^{|\text{miss}(i)|}$. Then $|\boldsymbol{\Sigma}_{\mathbf{X}}(\mathbf{y})|$ is quadratic and globally minimized in $\mathbf{y} = \boldsymbol{\mu}_{\mathbf{X} \text{ miss}(i)}(\mathbf{y}) + \boldsymbol{\Sigma}_{\mathbf{X} \text{ miss}(i), \text{obs}(i)}(\mathbf{y}) \boldsymbol{\Sigma}_{\mathbf{X} \text{ obs}(i), \text{obs}(i)}^{-1}(\mathbf{y}) ((\mathbf{x}_{i,1}, \dots, \mathbf{x}_{i,|\text{obs}(i)|}) - \boldsymbol{\mu}_{\mathbf{X} \text{ obs}(i)})$.*

Lemma 1 brings an additional insight: it allows to see the entire imputation (Algorithm 1 with Mahalanobis depth) as minimization of the covariance determinant. In this view, keeping all the points to be imputed but one fixed, covariance determinant is a quadratic function of the missing coordinates of this one point, and thus is minimized in a single point only. Thus, to impute points with missing coordinates one-by-one and iterate till convergence constitutes the block coordinate descent method, which proves to numerically converge due to Proposition 2.7.1 from Bertsekas (1999) (as long as $\boldsymbol{\Sigma}_{\mathbf{X}}$ is invertible).

Let $\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}} = \mathbf{U} \boldsymbol{\Lambda}^{\frac{1}{2}} \mathbf{V}'$ be the singular value decomposition (SVD) of the centered \mathbf{X} . Josse and Husson (2012) suggest the regularized PCA imputation, where, after an initialization, each point \mathbf{x} having missing values is imputed with \mathbf{y} such that $\mathbf{y}_j = \sum_{s=1}^S \mathbf{U}_{j,s} \sqrt{\frac{\lambda_s - \sigma^2}{\lambda_s}} \mathbf{V}_{j,s} + \boldsymbol{\mu}_{\mathbf{X} j}$ for all $j \in \text{miss}(\mathbf{x})$ and $\mathbf{y}_{\text{obs}(\mathbf{x})} = \mathbf{x}_{\text{obs}(\mathbf{x})}$ with $1 \leq S \leq d$ and some $0 < \sigma^2 \leq \frac{1}{d-S} \sum_{s=S+1}^d \lambda_s$; the algorithm proceeds iteratively till convergence. This method has proved its high efficiency in practice due to sticking to the low-rank structure of importance and ignoring noise. In addition, it can be seen as the truncated version of the iteratively applied extension of Stein’s estimator by Efron and Morris (1972).

We consider here its special case when $S = d$ and $0 < \sigma^2 \leq \lambda_d$. Proposition 2 is the regularized PCA analog to Proposition 1.

Proposition 2 (Imputation by regularized PCA reduces covariance determinant). *Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ be a sample in \mathbb{R}^d with $\mu_{\mathbf{X}} = 0$ and invertible $\Sigma_{\mathbf{X}}$. Further, for some $k \in \{1, \dots, n\}$, let $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ with $\mathbf{y}_i = \mathbf{x}_i$ for all $i \in \{1, \dots, n\} \setminus \{k\}$, $\mathbf{y}_{k,l} = \sum_{s=1}^d \mathbf{U}_{k,s} \sqrt{\frac{\lambda_s - \sigma^2}{\lambda_s}} \mathbf{V}_{l,s}$ with $0 < \sigma^2 \leq \lambda_d$ for all $i \in \text{miss}(k)$ and $\mathbf{y}_{k, \text{obs}(k)} = \mathbf{x}_{k, \text{obs}(k)}$, such that $\mathbf{y}_k \neq \mathbf{x}_k$. Then $|\Sigma_{\mathbf{Y}}| < |\Sigma_{\mathbf{X}}|$.*

Below, we state the main result of this section, that imputation using maximum Mahalanobis depth, iterative (multiple-output) regression, or regularized PCA with $S = d$ equilibrate at the same imputed sample.

Theorem 3. *Impute $\mathbf{X} = (\mathbf{X}_{\text{miss}}, \mathbf{X}_{\text{obs}})$ in \mathbb{R}^d with \mathbf{Y} so that for each $\mathbf{y}_i \in \mathbf{Y}$ with $|\text{miss}(i)| > 0$ holds $\mathbf{y}_{i, \text{miss}(i)} = \arg\max_{\mathbf{z}_{\text{obs}(i)} = \mathbf{y}_{\text{obs}(i)}} D^M(\mathbf{z} | \mathbf{Y})$. Then for each such \mathbf{y}_i holds as well:*

- \mathbf{x}_i is imputed with the **conditional mean**:

$$\mathbf{y}_{\text{miss}(i)} = \mu_{\mathbf{Y}_{\text{miss}(i)}} + \Sigma_{\mathbf{Y}_{\text{miss}(i), \text{obs}(i)}} \Sigma_{\mathbf{Y}_{\text{obs}(i), \text{obs}(i)}}^{-1} (\mathbf{x}_{\text{obs}(i)} - \mu_{\mathbf{Y}_{\text{obs}(i)}})$$

which is equivalent to **single- and multiple-output regression**,

- \mathbf{Y} is a **stationary point** of $|\Sigma_{\mathbf{X}}(\mathbf{X}_{\text{miss}})|$:

$$\frac{\partial |\Sigma_{\mathbf{X}}|}{\partial \mathbf{X}_{\text{miss}}}(\mathbf{Y}_{\text{miss}}) = \mathbf{0},$$

- each missing coordinate j of \mathbf{x}_i is imputed with **regularized PCA** by Josse & Husson (2012) with any $0 < \sigma^2 \leq \lambda_d$:

$$\mathbf{y}_{i,j} = \sum_{s=1}^d \mathbf{U}_{i,s} \sqrt{\frac{\lambda_s - \sigma^2}{\lambda_s}} \mathbf{V}_{j,s} + \mu_{\mathbf{Y}_j}.$$

Connection with the minimization of the covariance determinant expressed in the second point of Theorem 3 provides further insights. First, for a sample containing missing values, minimization of the covariance determinant corresponds to the minimization of the volume of the Mahalanobis depth lift $\text{mes}(D^M(X))$ defined as follows. Adding a real dimension to central regions $D_\alpha(X)$ and their multiplication with their depth α , $\alpha \in [0, 1]$, gives the depth lift

$$D(X) = \{(\alpha, \mathbf{y}) \in [0, 1] \times \mathbb{R}^d : \mathbf{y} = \alpha \mathbf{x}, \mathbf{x} \in D_\alpha(X), \alpha \in [0, 1]\}.$$

The depth lift is a body in \mathbb{R}^{d+1} , that describes location and scatter of the distribution of X , and in general gives rise to an ordering of distributions in \mathcal{M} (Mosler, 2013). Its specification for Mahalanobis depth is $D^M(X) = \{(\alpha, \mathbf{x}) \in [0, 1] \times \mathbb{R}^d : (\mathbf{x} - \alpha \mu_X)' \Sigma_X^{-1} (\mathbf{x} - \alpha \mu_X) \leq \alpha(1 - \alpha)\}$, and it possesses the volume

$$\text{mes}(D^M(X)) = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2} + 1)} \prod_{s=1}^{\lfloor \frac{d}{2} \rfloor} \frac{2s + 2(d \bmod 2)}{s + \frac{1}{2} + d \bmod 2} \left(1 - \left(1 - \frac{\pi}{8}\right) \mathbb{I}(d \bmod 2 \neq 0)\right) \sqrt{|\Sigma_X|}.$$

The connection between the Mahalanobis depth lift volume and the covariance determinant suggests an extension for the general depth function by imputing \mathbf{X}_{miss} with

$$\arg\min_{\mathbf{Y}_{\text{miss}} \in \mathbb{R}^{|\mathbf{X}_{\text{miss}}|}, \mathbf{Y}_{\text{obs}} = \mathbf{X}_{\text{obs}}} \text{mes}(D(\mathbf{Y})). \quad (6)$$

While being tractable for Mahalanobis depth, (6) demands enormous computational burden even for very moderate data sets, as a single evaluation of $\text{mes}(D(\mathbf{Y}))$ amounts to $\text{mes}(D^{z(n)}(\mathbf{X})) = \frac{1}{n^{d+1}} \sum_{\{i_0, \dots, i_d\} \subset \{1, \dots, n\}} \left| ((1, \mathbf{x}'_{i_0})', \dots, (1, \mathbf{x}'_{i_d})') \right|$ for zonoid depth and to $\text{mes}(D^{T(n)}(\mathbf{X})) = \sum_{i=1}^{n\alpha_{max}} \frac{i^{d+1} - (i-1)^{d+1}}{(d+1)n^{d+1}} \text{mes}(D_{\frac{i}{n}}^{T(n)}(\mathbf{X}))$ for Tukey depth with α_{max} being the depth of the Tukey median w.r.t. the sample \mathbf{X} . On the other hand, it constitutes a separate approach giving a solution different to the proposed one. For these reasons we leave it out of the scope of the present article.

5 Experimental study

5.1 Choice of competitors

The proposed methodology is represented by the three introduced imputation schemes based on the iterative maximization of Tukey, zonoid, and Mahalanobis depth. Further, we include the same imputation scheme using Mahalanobis depth based on MCD mean and covariance estimates with robustness parameter chosen in an optimal way due to knowledge of the simulation setting. Next, conditional mean imputation based on EM estimates of mean and covariance matrix is taken. We also include two regularized PCA imputations assuming the data rank to be equal to 1 and to 2. Further, two nonparametric imputation methods are used, namely random forest with its default implementation in R-package `missForest` and k NN imputation tuned as described in Algorithm 2 (Section 3) by [Stekhoven and Bühlmann \(2012\)](#), *i.e.* by choosing k from $\{1, \dots, 15\}$ minimizing imputation error over 10 validation sets. Finally, for the benchmark purposes, mean and oracle (if possible) imputations are added.

5.2 Simulated data

We start by exploring the MCAR mechanism applied to the family of elliptically symmetric Student- t distributions. Regarding Definition 1, we set center $\boldsymbol{\mu}_2 = (1, 1, 1)'$, shape $\boldsymbol{\Sigma}_2 = ((1, 1, 1)', (1, 4, 4)', (1, 4, 8)')$, and let F be the univariate Student- t distribution ranging the number of degrees of freedom (d.f.) from the Gaussian to the Cauchy: $t = \infty, 10, 5, 3, 2, 1$. For each of the 1000 random simulations, we remove 25% of values due to MCAR, and indicate the median and the median absolute deviation from the median (MAD, in parentheses) of the RMSE of each imputation method for a sample of size 100 points in Table 1. In each column of Table 1, we distinguish the best method in bold and the second best in italics, ignoring the oracle imputation. We set robustness parameter of the MCD to 0.75, an optimal choice after several tries; also because imputed points concentrate in subspaces of lower dimension and this singularity hinders execution of the MCD algorithm with lower parameter values already on the stage of initialization. The oracle imputes with the conditional mean using population parameters $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$.

For the presented range of elliptical Student- t distributions, behavior of different imputation methods changes with the number of d.f., as well as the general tendency of the leadership. For the Cauchy distribution, robust methods perform best: the Mahalanobis depth-based imputation using MCD estimates, closely followed by the one using Tukey depth. For 2 d.f., when the first moment exists but the second does not, EM- and Tukey-depth-based imputation perform similarly, with a slight advantage of the Tukey depth in terms of the MAD. For larger numbers of d.f., when two first moments exist, EM takes the leadership. It is followed by the group of the

Distr.	Gaussian	t_{10}	t_5	t_3	t_2	Cauchy
D^{Tuk}	1.675 (0.205)	1.871 (0.2445)	2.143 (0.3313)	2.636 (0.5775)	3.563 (1.09)	16.58 (13.71)
D^{zon}	1.609 (0.1893)	1.81 (0.2395)	2.089 (0.3331)	2.603 (0.5774)	3.73 (1.236)	19.48 (16.03)
D^{Mah}	1.613 (0.1851)	1.801 (0.2439)	2.079 (0.3306)	2.62 (0.5745)	3.738 (1.183)	19.64 (16.2)
$D_{MCD.75}^{Mah}$	1.991 (0.291)	2.214 (0.3467)	2.462 (0.4323)	2.946 (0.6575)	3.989 (1.287)	16.03 (12.4)
EM	1.575 (0.1766)	1.755 (0.2379)	2.026 (0.3144)	2.516 (0.5537)	3.567 (1.146)	18.5 (15.46)
regPCA1	1.65 (0.1846)	1.836 (0.2512)	2.108 (0.3431)	2.593 (0.561)	3.692 (1.186)	18.22 (15.02)
regPCA2	1.613 (0.1856)	1.801 (0.2433)	2.08 (0.3307)	2.619 (0.5741)	3.738 (1.19)	19.61 (16.1)
k NN	1.732 (0.2066)	1.923 (0.2647)	2.235 (0.3812)	2.757 (0.5874)	3.798 (1.133)	17.59 (14.59)
RF	1.763 (0.2101)	1.96 (0.2759)	2.259 (0.3656)	2.79 (0.5856)	3.849 (1.19)	17.48 (14.33)
mean	2.053 (0.2345)	2.292 (0.2936)	2.612 (0.3896)	3.165 (0.6042)	4.341 (1.252)	20.32 (16.36)
oracle	1.536 (0.1772)	1.703 (0.2206)	1.949 (0.3044)	2.384 (0.5214)	3.175 (0.9555)	13.55 (10.71)

Table 1: Median and MAD of the RMSE of the imputation for a sample of 100 points drawn from the family of elliptically symmetric Student- t distributions with parameters μ_2 and Σ_2 having 25% of missing values due to MCAR, over 1000 repetitions.

regularized PCA methods, and Mahalanobis- and zonoid-depth-based imputation. Please note that Mahalanobis-depth and regularized PCA with two-dimensional low-rank model perform in the same way (the tiny difference can be explained by the precision constant), see Theorem 3. Both nonparametric imputation methods perform rather poorly being “unaware” of the ellipticity of the underlying distribution, and deliver (to a certain extent) reasonable results for the case of the Cauchy distribution only where partial insensibility to the correlation between the variables can be seen as an advantage.

In the second simulation, we modify the above setting by adding 15% of outliers that stem from Cauchy distribution with the same parameters μ_2 and Σ_2 . The portion of missing values is kept on the same level of 25% but the outlying observations do not contain missing values. The parameter of the MCD algorithm for the robust Mahalanobis-depth-based imputation is set to 0.85, *i.e.* exactly corresponding to the portion of non-contaminated data. Corresponding medians and MADs of the RMSE over 1000 repetitions are indicated in Table 2.

As expected, best RMSEs are obtained by the robust imputation methods: Tukey depth and Mahalanobis depth with MCD estimates. Being restricted to a neighborhood, nonparametric methods often impute based on non-outlying points, and thus deliver second best imputation group. The rest of the included imputation methods do not resist pollution of the data and perform rather poorly.

Further, we explore the performance of the proposed methodology in a MAR setting. For this, first, we generate highly correlated Gaussian data by setting mean to $\mu_3 = (1, 1, 1)$ and

Distr.	Gaussian	t_{10}	t_5	t_3	t_2	Cauchy
D^{Tuk}	1.751 (0.2317)	1.942 (0.2976)	2.178 (0.3556)	2.635 (0.6029)	3.763 (1.17)	17.17 (13.27)
D^{zon}	1.86 (0.3181)	2.087 (0.4295)	2.333 (0.4924)	2.864 (0.7819)	4.082 (1.535)	20.43 (15.99)
D^{Mah}	1.945 (0.4299)	2.165 (0.5473)	2.421 (0.6026)	2.935 (0.8393)	4.136 (1.501)	20.27 (15.91)
$D_{MCD.85}^{Mah}$	1.81 (0.239)	2.022 (0.3128)	2.231 (0.381)	2.664 (0.5877)	3.783 (1.224)	16.46 (12.94)
EM	1.896 (0.3987)	2.112 (0.5226)	2.376 (0.5715)	2.828 (0.7773)	4.036 (1.518)	19.01 (15.21)
regPCA1	1.958 (0.4495)	2.196 (0.5729)	2.398 (0.6035)	2.916 (0.8221)	4.09 (1.585)	19.81 (16.15)
regPCA2	1.945 (0.4328)	2.165 (0.5479)	2.421 (0.5985)	2.93 (0.8384)	4.14 (1.503)	20.53 (16.28)
k NN	1.859 (0.2602)	2.051 (0.3143)	2.315 (0.3809)	2.797 (0.6045)	3.955 (1.265)	18.96 (14.73)
RF	1.86 (0.2332)	2.047 (0.3043)	2.325 (0.3946)	2.838 (0.6228)	4.026 (1.354)	19.04 (14.62)
mean	2.23 (0.3304)	2.48 (0.4163)	2.766 (0.528)	3.34 (0.7721)	4.623 (1.561)	21.04 (15.56)
oracle	1.563 (0.1849)	1.733 (0.2266)	1.939 (0.2979)	2.356 (0.4946)	3.323 (1.04)	14.44 (11.33)

Table 2: Median and MAD of the RMSE of the imputation for a sample of 100 points drawn from the family of elliptically symmetric Student- t distributions with parameters μ_2 and Σ_2 contaminated with 15% of outliers having 25% of missing values due to MCAR on non-contaminated data, over 1000 repetitions.

covariance matrix to $\Sigma_3 = ((1, 1.75, 2)', (1.75, 4, 4)', (2, 4, 8)')$. Second, we introduce missing values depending on the existing values according to the following scheme: first variable has missing value with probability 0.08 if second variable is higher than the population mean, and with probability 0.7 if second variable is lower than the population mean; for the third variable corresponding probabilities constitute 0.48 and 0.24. This mechanism leads to a highly asymmetric pattern of 24% MAR values. The boxplots of the RMSEs of the considered imputation methods over 1000 repetitions are indicated in Figure 5.

According to our expectations, semi-parametric methods (EM- and regularized-PCA-based imputation, and thus Mahalanobis-depth-based imputation as well) perform well and close to the oracle imputation. Better performance when considering the one-dimensional low-rank model for the regularized PCA can be explained by the high correlation. Though having no parametric knowledge, zonoid-depth-based imputation also performs satisfactorily. Being unable to capture sufficiently the correlation, nonparametric methods perform poorly. Even worse performance of robust methods is explained by the fact of “throwing away” points that possibly contain valuable information.

Finally, we consider an extremely contaminated low-rank model. Namely, we fix a two-dimensional low-rank structure and add Cauchy-distributed noise. Then, we remove 20% of values according to the MCAR mechanism. The resulting medians and MADs of the RMSE over 1000 repetitions are indicated in Table 3. We exclude the oracle imputation, as it is supposed to

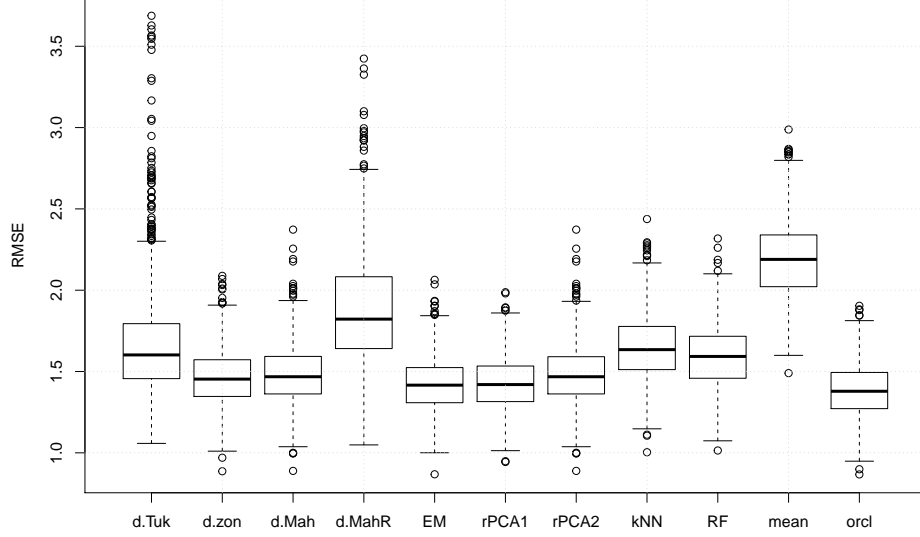


Figure 5: Root means square errors for different imputation methods for a correlated three-dimensional normal sample with parameters μ_3 and Σ_3 of 100 points with missing values according to MAR, over 1000 repetitions.

be always correct and thus cannot serve as a benchmark. This model can be seen as a stress-test of the regarded imputation methods. In general, as capturing any structure is rather meaningless in this setting (which is additionally confirmed by the high MADs), the performance of the methods is “proportional to the way they ignore” dependency information. For this reason, mean imputation as well as nonparametric methods perform best. On the other hand, one should note that accounting only for fundamental features of the data, Tukey-depth- and zonoid-depth-based methods perform second best. This can be also said about the regularized PCA keeping the first principal component only. The rest of the methods try to reconstruct the data structure, and being distracted either by the low rank or by the heavy-tailed noise show poor results.

	D^{Tuk}	D^{zon}	D^{Mah}	$D_{0.75}^{MahR}$	EM	regPCA1	regPCA2	kNN	RF	mean
Median RMSE	0.4511	0.4536	0.4795	0.5621	0.4709	0.4533	0.4664	0.4409	0.4444	<i>0.4430</i>
Mad of RMSE	0.3313	0.3411	0.3628	0.4355	0.3595	0.3461	0.3554	0.3302	0.3389	<i>0.3307</i>

Table 3: Medians and MADs of RMSE for a two-dimensional low-rank model in \mathbb{R}^4 of 50 points with the Cauchy-distributed noise and 20% of missing values according to MCAR, over 1000 repetitions.

5.3 Real data

In addition, we validate the proposed methodology on four real benchmark data sets taken from the UCI Machine Learning Repository (<http://archive.ics.uci.edu/ml>). For each data set, we consider part of variables and observations only (picking one class) in order to exclude categorical variables and reduce computational burden. The finally handled data sets are: Banknotes ($n = 100$, $d = 3$), Glass ($n = 76$, $d = 3$), Pima ($n = 68$, $d = 4$), Blood Transfusion ($n = 502$,

$d = 3$, Yeh et al., 2009). Full details on the experimental design are given in the reproducing sources at <https://github.com/julierennes>. We investigate the same set of imputation methods, set parameter of the MCD to 85%, and exclude oracle again as it would produce perfect imputation. The RMSE's boxplots over 500 repetitions of imputation after removing 15% of entries due to MCAR mechanism are depicted in Figure 6. First, we should note that throughout the data sets, zonoid-depth-based imputation stably delivers highly satisfactory results.

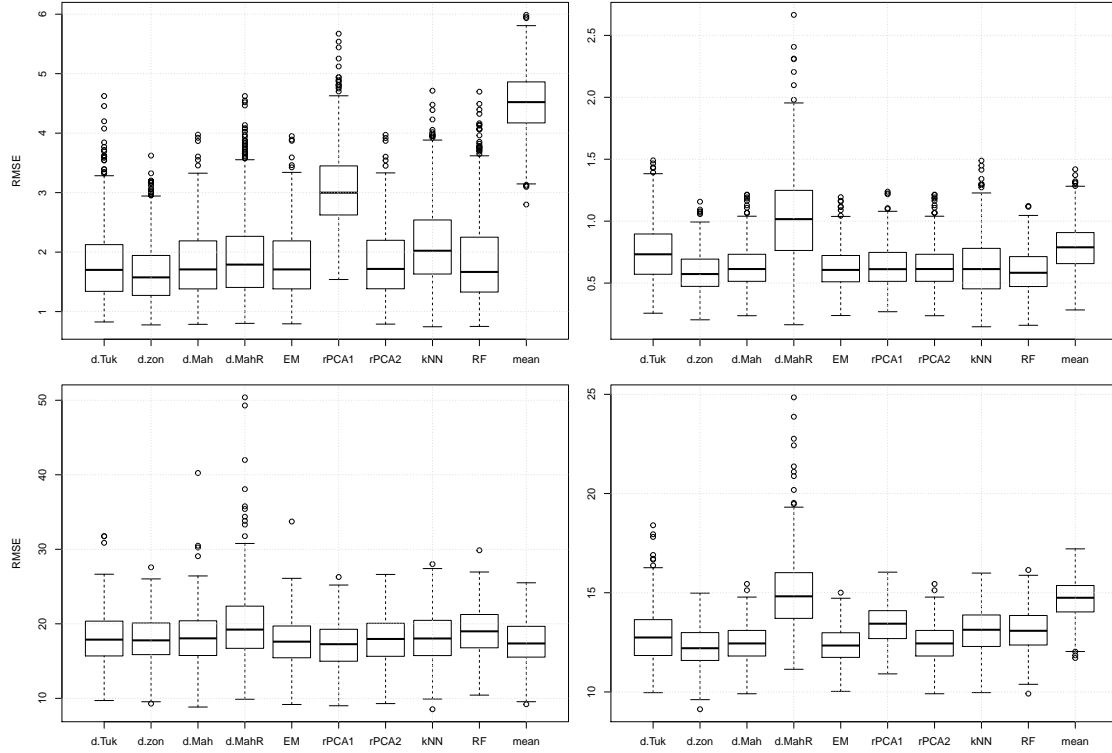


Figure 6: RMSEs for Banknotes (top, left), Glass (top, right), Pima (bottom, left), and Blood Transfusion (bottom, right) data sets with 15% of MCAR values over 500 repetitions.

Visual inspection of the Banknotes data set shows that it rather constitutes a mixture of two components, and thus mean imputation as well as the one-dimensional regularized PCA show poor performance. Random forest imputation delivers satisfactory results by fitting to the local geometry of the data. On the other hand, zonoid-depth-based imputation searching compromise between local and global features delivers best results. Unsatisfactory performance of the k NN imputation could be explained by local inhomogeneities in the data (points form plenty of local clusters of different size), *i.e.* it is too local; this problem seems to be captured by the aggregation stage of the random forest. All methods imputing by conditional mean (both Mahalanobis-depth-based, EM-based, and regularized PCA imputation) perform reasonably as well, while imputing in two-dimensional affine subspaces. Tukey-depth-based imputation captures geometry of the data on one side but lacks information for robustness reasons (outliers do not seem to be a problem here), and thus performs similar.

Glass data turns out to be more challenging as it highly deviates from ellipticity and part of the data lie sparse in the space but do not seem to be outlying (perhaps a separate component

containing among others limit or censored cases). Thus, mean and robust Mahalanobis-depth- and Tukey-depth-based imputations perform poorly. An additional disadvantage for the Tukey depth is presence of a large number of outsiders, which are imputed with maximum spatial depth using obviously non-optimal here MCD estimates. For the same reason, both (groups) of semi- and non-parametric methods do not outperform mean that much. Accounting for local geometry, random forest and zonoid-depth-based imputation perform slightly better, and deliver best results.

Even more challenging is the Pima data set. Its variables are only weakly dependent on each other, some correlation can be observed between third and fourth variables, but even this is weakened by the presence of an outlier in third dimension, which lies on the mean level in dimensions two and four. Similar to the highly contaminated low-rank model from Section 5.2, in this setting partial ignorance regarding dependency is of advantage. For this reason, mean and one-dimensional regularized PCA perform best, but depth-based imputation methods are included in the closely following group. Methods accounting for data ellipticity perform comparably because they are able to provide close to reasonable imputation at least in the two correlated dimensions.

Blood Transfusion data visually reminds a tetrahedron being dispersed away from one of its vertices. Thus, mean imputation can be substantially improved. As nonparametric methods disregard dependency between dimension and one-dimensional regularized PCA does not capture this sufficiently, they perform poorly. Better imputation is delivered by the depth- and EM-based methods, those capturing correlation. As data still deviates from the ellipticity and a few outliers are present only, MCD throwing away 15% of the data worsen robust Mahalanobis-depth-based imputation; this effect is partially transferred to the Tukey-depth-based imputation, via the outsider treatment. Due to the same reason of deviation from ellipticity, zonoid-depth-based imputation is further slightly advantageous.

6 Multiple imputation for elliptical family

The developed above generic framework allows to go beyond the single imputation and enable for statistical inference by means of multiple imputation (Little and Rubin, 2002). This approach consists in calculating estimator of interest on a number of generated imputed data sets with further aggregation. Drawing multiply imputed data sets is traditionally performed in two steps: The first step consists in reflecting the uncertainty of the parameters of the imputation model. The second step consists in imputing close to the underlying distribution. Imputation model uncertainty may be reflected using either bootstrap or Bayesian approach, see e.g. Schafer (1997), Efron (1994), and for multivariate normal setting EM-estimates can be used to draw from the conditional normal distribution. An alternative is to employ the Markov chain Monte Carlo (MCMC) with normal conditional distributions, viz. multiple imputation by chained equations (MICE) by van Buuren (2012).

Depth-based single-imputation framework introduced above allows to extend multiple imputation in a natural way to the more general elliptical setting. We start by showing how to reflect uncertainty due to distribution with that delivers so-called improper imputation (Section 6.1). After, using bootstrap to reflect model uncertainty, we state the complete algorithm in Section 6.2.

6.1 Accounting for uncertainty due to distribution

When imputing an observation x in case of multivariate normality, one can use EM estimates to draw x_{miss} from $N(\mu, \Sigma)$ conditioned on x_{obs} , for instance using the Schur complement

(stochastic EM). Conditional distribution of an elliptically symmetric distribution (we assume absolute continuity and that center and shape can be consistently estimated as $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$) is elliptical as well, and can be derived by the proper transformation of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and the univariate (radial) density (Fang et al., 1990). Estimation of these preserves two complications: incompleteness of the data and one-sided possibly heavy-tailed density. To overcome the first one, we design a MCMC allowing to use estimators on complete data. For the second issue, taking into account that multiple imputation rather requires drawing a point then estimating the density itself, we stay in the depth framework and proceed as follows. First, using depth c.d.f. we draw the depth contour. And second, using $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, we draw the point containing missing values in the subspace of the intersection of this contour with the hyperplane of missing coordinates. The suggested procedure allows us to draw from the conditional distribution in a semiparametric way, as detailed explicitly right below.

For an absolutely continuous elliptically symmetric distribution, any data depth satisfying corresponding postulates from Mosler (2013) or Zuo and Serfling (2000) possesses the characterization property: it is a monotone function of the radial density, which in order is expected to be a monotone function of the Mahalanobis distance. For the Mahalanobis depth for instance, which is just a monotone transformation of the Mahalanobis distance, this relationship is the most intuitive.

Different to the normal case, the shape of each conditional distribution depends on the condition itself. Precisely, this heteroscedasticity is determined by Mahalanobis distance of the point \mathbf{x} (to be imputed) and of the conditional center $\boldsymbol{\mu}^*$ from the unconditional one $\boldsymbol{\mu}$. Given a complete data set \mathbf{X} , $\boldsymbol{\mu}^*$ can be obtained due to (3). Further, let $f_{D(\mathbf{X}|\mathbf{X})}$ denote the density of the depth for a random vector $\mathbf{X} \in \mathbb{R}^d$ w.r.t. itself. The depth of the imputing point \mathbf{y} should then be drawn as a quantile Q uniformly on $[0, F_{\boldsymbol{\mu}^*}(D(\boldsymbol{\mu}^*|\mathbf{X}))]$, where

$$F_{\boldsymbol{\mu}^*}(x) = \int_0^x f_{D(\mathbf{X}|\mathbf{X})}(z) \frac{\left(\sqrt{d_{Mah.}^2(z) - d_{Mah.}^2(D(\boldsymbol{\mu}^*|\mathbf{X}))} \right)^{|miss(\mathbf{x})|-1}}{d_{Mah.}^{d-1}(z)} \times \quad (7)$$

$$\times \frac{d_{Mah.}(z)}{\sqrt{d_{Mah.}^2(z) - d_{Mah.}^2(D(\boldsymbol{\mu}^*|\mathbf{X}))}} dz,$$

with $d_{Mah.}(z)$ being the Mahalanobis distance to the center as a function of depth. Then, Q is simply projected back on the support by $\alpha = F_{\boldsymbol{\mu}^*}^{-1}(Q)$ corresponding to the surface of $D_\alpha(\mathbf{X})$ — the trimmed region of depth α , see Figure 8 (left). The aim of transformation (7) is to normalize the volume (see Appendix for the derivation). Any constant normalization factor can be omitted here as $F_{\boldsymbol{\mu}^*}$ is used exceptionally for drawing. The square root in the formula could be shortened, but this way Mahalanobis distance is exploited as a function of depth for joint distribution only, which can be estimated from the data without further transformations. For instance, when using the Mahalanobis depth, one can substitute directly in the equation (7) $d_{Mah.}(y)$ by $\sqrt{1/y - 1}$.

Next, the point $\mathbf{y} \in \partial D_\alpha(\mathbf{X}) \cap \{\mathbf{z} \in \mathbb{R}^d \mid \mathbf{z}_{obs(\mathbf{x})} = \mathbf{x}_{obs}\}$ (lying in intersection of the region of depth α with the hyperplane of missing values of \mathbf{x}) should be randomly chosen. This is done by drawing \mathbf{u} uniformly on $\mathcal{S}^{|miss(\mathbf{x})|-1}$ and transforming it by conditional scatter matrix obtaining $\mathbf{u}^* \in \mathbb{R}^d$ having $\mathbf{u}_{miss(\mathbf{x})}^* = \boldsymbol{\Lambda} \mathbf{u}$ (with $\boldsymbol{\Lambda}(\boldsymbol{\Lambda})' = \boldsymbol{\Sigma}_{miss(\mathbf{x}),miss(\mathbf{x})} - \boldsymbol{\Sigma}_{miss(\mathbf{x}),obs(\mathbf{x})} \boldsymbol{\Sigma}_{obs(\mathbf{x}),obs(\mathbf{x})}^{-1} \boldsymbol{\Sigma}_{obs(\mathbf{x}),miss(\mathbf{x})}$) and $\mathbf{u}_{obs(\mathbf{x})}^* = \mathbf{0}$. Such \mathbf{u}^* is uniformly distributed on the conditional depth contour. Then \mathbf{x} is imputed as $\mathbf{y} = \boldsymbol{\mu}^* + \beta \mathbf{u}^*$, where β is a scalar obtained as the positive solution of $\boldsymbol{\mu}^* + \beta \mathbf{u}^* \in \partial D_\alpha(\mathbf{X})$ (e.g. quadratic equation

$(\mu^* + \beta u^* - \mu)' \Sigma^{-1} (\mu^* + \beta u^* - \mu) = d_{Mah.}^2(\alpha)$ in the case of Mahalanobis depth). See Figure 8, right for an illustration.

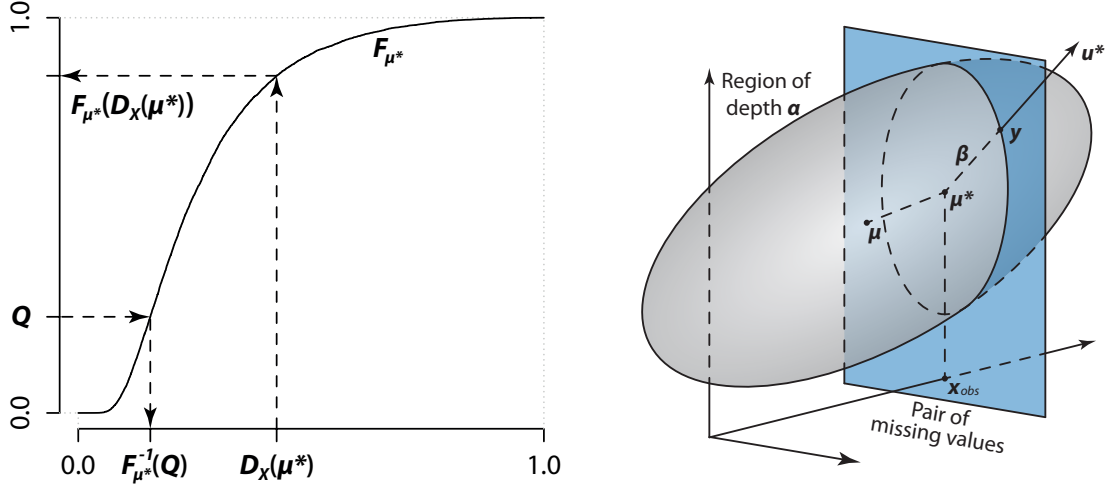


Figure 7: Illustration of an application of (7) to impute by drawing from the conditional distribution of an elliptical distribution. Drawing depth $D = F_{\mu^*}^{-1}(Q)$ via the depth c.d.f. F_{μ^*} (left) and locating the corresponding imputed point y (right).

Proposed method allows to impute by conditional drawing from an elliptically symmetric absolutely continuous distribution, and can be simply employed for flat ones generalizing the scatter-matrix operations to the singular cases. We shortly demonstrate its capabilities by the following simulation study. We generate 500 points from an elliptical Student- t distribution with 3 degrees of freedom with mean $\mu = (-1, -1, -1, -1)'$ and structure matrix $((0.5, 0.5, 1, 1)', (0.5, 1, 1, 1)', (1, 1, 4, 4)', (1, 1, 4, 10)')$, put 30% of missing values due to MCAR, and impute them reflecting uncertainty due to distribution only by stochastic EM and the proposed method. Medians of the univariate quantiles over 2000 repetitions for the initial distribution, stochastic EM, and suggested algorithm are indicated in Table 4. While stochastic EM, generating noise from the normal model, fails to guess the quantiles as expected, proposed method deliver their quite precise exploration with only slight deviations in the tail of the distribution due to (again expected) complication to reflect density shape there. Though such an output is expected, this greatly extends the scope of practices compared to the deep-rooted Gaussian-based imputation.

6.2 Inference for incomplete data

Imputation scheme designed in Section 6.1 accounts for uncertainty w.r.t. distribution only, and thus cannot be directly used for deriving inference from incomplete data. To perform proper multiple imputation, we resort to the bootstrap approach to reflect uncertainty due to the estimation of the underlying semi-parametric model as well: To generate each imputed data set, we first draw a sequence of indices $b = (b_1, \dots, b_n)$ with $b_i \sim U(1, \dots, n)$ for $i = 1, \dots, n$, and then utilize this sequence to obtain a subset (with repetitions) $X_{b,\cdot} = (x_{b_1}, \dots, x_{b_n})$ used to perform single imputation giving μ^* and to estimate the shape Σ on each Monte-Carlo iteration. The depth-based procedure for multiple imputation (called DMI) can be described by Algorithm 2. Taking single

Quantile:	0.5	0.75	0.85	0.9	0.95	0.975	0.99	0.995
X_1 :								
complete	-1.0013	-0.4632	-0.1231	0.1446	0.6398	1.2017	2.0661	2.8253
stoch. EM	-1.0008	-0.4225	-0.0649	0.2114	0.6902	1.2022	1.9782	2.6593
depth	-0.9996	-0.4643	-0.1232	0.1491	0.6509	1.2142	2.0827	2.8965
X_2 :								
complete	-0.9992	-0.2359	0.2416	0.6237	1.3252	2.1263	3.3277	4.3852
stoch. EM	-1.0018	-0.1675	0.3468	0.7468	1.4205	2.1355	3.1723	4.1049
depth	-1.0008	-0.2318	0.2522	0.6411	1.3537	2.1627	3.3771	4.5018
X_3 :								
complete	-1.0030	0.5194	1.4815	2.2513	3.6353	5.2170	7.6330	9.8235
stoch. EM	-1.0038	0.6304	1.6537	2.4386	3.7903	5.2302	7.4294	9.2877
depth	-1.0043	0.5228	1.4919	2.2660	3.6624	5.2602	7.7358	9.8991
X_4 :								
complete	-1.0134	1.3938	2.9121	4.1352	6.3271	8.8592	12.6258	16.1014
stoch. EM	-0.9968	1.6388	3.2870	4.5494	6.6477	8.8909	12.0789	14.9915
depth	-1.0091	1.3986	2.9341	4.1559	6.3862	8.9471	12.7642	16.3143

Table 4: Median univariate quantiles of imputed elliptical sample consisting of 500 points drawn from Student- t distribution with 3 d.f. over 2000 repetitions.

imputation as a starting point allows to begin Markov chain Monte-Carlo closer to the stationary mode and thus reduce the burn-in period to a few iterations only.

Algorithm 2 Depth-based multiple imputation

```

1: function IMPUTE.DEPTH.MULTIPLE( $\mathbf{X}$ , num.burnin, num.sets)
2:   for  $m = 1 : \text{num.sets}$  do
3:      $\mathbf{Y}^{(m)} \leftarrow \text{IMPUTE.DEPTH.SINGLE}(\mathbf{X})$   $\triangleright$  Start MCMC with a single
       imputation
4:      $\mathbf{b} \leftarrow (b_1, \dots, b_n) = (U(1, \dots, n), \dots, U(1, \dots, n))$   $\triangleright$  Draw bootstrap sequence
5:     for  $k = 1 : (\text{num.burnin} + 1)$  do
6:        $\Sigma \leftarrow \hat{\Sigma}(\mathbf{Y}_{\mathbf{b}, \cdot}^{(m)})$ 
7:       Estimate  $f_{D(X|X)}$  using  $\mathbf{Y}^{(m)}$ .
8:       for  $i = 1 : n$  do
9:         if  $\text{miss}(i) \neq \emptyset$  then
10:           $\mu^* \leftarrow \text{IMPUTE.DEPTH.SINGLE}(\mathbf{x}_i, \mathbf{Y}_{\mathbf{b}, \cdot}^{(m)})$   $\triangleright$  Single-impute point
11:           $\mathbf{u} \leftarrow U(\mathcal{S}^{|\text{miss}(i)|-1})$ 
12:           $\mathbf{u}_{\text{miss}(i)}^* \leftarrow \mathbf{u}\mathbf{\Lambda}$   $\triangleright$  Calculate random direction
13:           $\mathbf{u}_{\text{obs}(i)}^* \leftarrow 0$ 
14:          Calculate  $F_{\mu^*}$ 
15:           $Q \leftarrow U([0, F_{\mu^*}(D(\mu^*))])$   $\triangleright$  Draw depth
16:           $\alpha \leftarrow F_{\mu^*}^{-1}(Q)$ 
17:           $\beta \leftarrow$  positive solution of  $\mu^* + \beta \mathbf{u}^* \in \partial D_\alpha(\mathbf{Y}_{\mathbf{b}, \cdot}^{(m)})$ .
18:           $\mathbf{y}_{i, \text{miss}(i)}^{(m)} \leftarrow \mu_{\text{miss}(i)}^* + \beta \mathbf{u}_{\text{miss}(i)}^*$   $\triangleright$  Impute one point
19:   return  $(\mathbf{Y}^{(1)}, \dots, \mathbf{Y}^{(\text{num.sets})})$ 

```

The function `impute.depth.single`($\mathbf{x}_i, \mathbf{Y}_{\mathbf{b}, \cdot}^{(m)}$) in Algorithm 2 corresponds to the obvious modification of Algorithm 1 for imputing missing values of single point \mathbf{x}_i calculating all the depth estimates w.r.t. the complete data set $\mathbf{Y}_{\mathbf{b}, \cdot}^{(m)}$. $\mathbf{\Lambda}$ is such that $\mathbf{\Lambda}(\mathbf{\Lambda})' = \Sigma$. Depth density $f_{D(X|X)}$ can be obtained using any consistent estimate, while for F_{μ^*} numerical integration can be employed. In general, Algorithm 2 sticks to the above notation, and returns a number of multiply-imputed data sets.

While estimation of the depth density gives a clear advantage for DMI over existing implementations since it captures the joint distribution, its competitiveness under the Gaussian setting remains interesting. Thus, we explore the performance of DMI in estimating coefficients of two regression model. The first one is: $Y = \beta'(1, X')' + \epsilon$ where $\beta = (0.5, 1)'$, $X \sim N(1, 4)$, and $\epsilon \sim N(0, 0.25)$. 30% of missing values are introduced due to MCAR. We employ DMI and add multiple imputation by R-packages `Amelia` and `mice` (with Gaussian conditional distribution) with the default settings for comparison. We generate 5 and 20 multiply-imputed data sets. Based on a sample consisting of 100 observations, over 1000 repetitions, we indicate medians, coverage by the 95% confidence interval constructed according to the Rubin's rule, and width of this interval of the estimates of β in Table 5.

Note that both `Amelia` and `mice` are optimal in these settings. In the experiment, `Amelia` seems to slightly undercover, more expressed when considering 20 multiply-imputed data sets. `mice` delivers very reasonable estimates. DMI suffers from a slight overcovering.

	β_0			β_1		
	med	cov	width	med	cov	width
5 multiply-imputed data sets						
Amelia	0.498	0.941	0.39	1.004	0.947	0.168
mice	0.506	0.945	0.39	1	0.943	0.178
DMI	0.501	0.966	0.484	0.998	0.971	0.212
20 multiply-imputed data sets						
Amelia	0.496	0.937	0.351	1	0.931	0.156
mice	0.499	0.957	0.371	0.997	0.946	0.166
DMI	0.5	0.979	0.421	0.996	0.969	0.191

Table 5: Medians (med), 95% coverage due to the Rubin’s rule (cov), and width of the confidence intervals (width) for the first regression model sample consisting of 100 observations with 30% MCAR-coordinates, based on 5 and 20 multiply imputed data sets, over 1000 repetitions.

The second regression model is $Y = \beta'(1, X')' + \epsilon$ with $\beta = (0.5, 1, 3)'$ and $X \sim N((1, 1)', ((1, 1)', (1, 4)'))$. Resting settings are kept. Except for missing values, the joint elliptical distribution of $(X', Y)'$ reserves additional difficulties due to high correlation (≈ 0.988) between the second component of X and Y . We indicate the results in Table 6.

	β_0			β_1			β_2		
	med	cov	width	med	cov	width	med	cov	width
5 multiply-imputed data sets									
Amelia	0.5	0.946	0.536	1.005	0.939	0.438	2.999	0.94	0.226
mice	0.525	0.984	1.464	1.063	0.975	1.476	2.92	0.976	0.88
DMI	0.513	0.974	0.719	0.989	0.957	0.589	3	0.961	0.295
20 multiply-imputed data sets									
Amelia	0.487	0.931	0.489	1.01	0.941	0.399	2.998	0.929	0.206
mice	0.519	0.984	1.6	1.081	0.98	1.807	2.881	0.982	1.502
DMI	0.504	0.971	0.613	0.989	0.979	0.519	3.003	0.97	0.26

Table 6: Medians (med), 95% coverage due to the Rubin’s rule (cov), and width of the confidence intervals (width) for the second regression model sample consisting of 100 observations with 30% MCAR-coordinates, based on 5 and 20 multiply imputed data sets, over 1000 repetitions.

Amelia still slightly undercovers, again more pronounced for 20 multiply-imputed data sets. mice delivers biased coefficients, due to its MCMC-nature and high correlation that causes instability in regression models, and seriously overcovers (even visually its 95% confidence intervals a substantially larger). DMI on the other hand, retains the same behavior suffering from a slight overcovering, but in general delivers reasonable results. It is worth to notice that due to the estimation of the depth density the application of the Rubin’s rule to DMI is not theoretically justified and requires further investigation (Reiter and Raghunathan, 2007). This could explain the little overcovering, which still keeps the confidence level and is better than undercovering.

7 Conclusions

The proposed framework for imputation based on data depth fills the gap between global imputation pursued by regression- and PCA-based methods, and the local one represented by *e.g.* random forest or k NN imputation. It reflects unsureness in the distribution assumption by imputing close to data geometry, is robust in sense of distribution and outliers, and preserves functionality under MAR mechanism. When used with Mahalanobis depth, exploiting data depth as a concept the connection between iterative regression, regularized PCA, and minimum covariance determinant imputation has been established. Empirical study shows efficiency of the suggested methodology for various elliptical distributions and elliptically resembling real data settings. Further, in a natural way the method is extended to multiple imputation for the elliptical family, which enlarges the area of application of the existing tools for multiple imputation.

The methodology is generic, *i.e.* any reasonable notion of data depth can be employed, which will determine the properties of the imputation. Due to empirical study, zonoid depth behaves well in general, and in the real-data settings particularly. On the other hand Tukey depth may be preferred if robustness is an issue. Further, projection depth (Zuo and Serfling, 2000) is a proper choice if only a few points contain missing values in a data set that is substantially outlier-contaminated. This specific case is not included in the article but the projection-depth-based imputation can be found in the implementation. To reflect multimodality of the data, the suggested framework can be employed with localized depths, see *e.g.* Paindaveine and Bever (2013), whereas the localization parameter can be tuned by means of cross-validation due to the imputation quality (*e.g.* as it was done for tuning the k NN imputation in Section 5).

A serious question with data depths is their computation demand. For the purity of experiment, in Section 5 all the computations were using exact algorithms, that is why the study is restricted to dimensions 3 and 4 and a few hundred points only. Even using approximate versions of data depths (which can be found in the implementation as well) does not scale the amount of data that can be handled substantially. Imputing large data sets might be overcome by the expected developments in the field of data depth.

The methodology has been implemented as an R-package. Source codes of the package and of the experiment-reproducing files can be downloaded from <https://github.com/julierennes>.

Appendix: Proofs

Proof of Theorem 1:

The proof for Mahalanobis depth is obvious. Exploiting the results by Liu and Singh (1993), for zonoid and Tukey depths one obtains:

For zonoid depth, due to absolute continuity of \mathcal{E} , $\exists N_{ch}$ such that $\mathbf{x} \in \text{conv}(\mathbf{X}_{obs}^{(n)}) \forall n > N_{ch}$, and thus due to continuity of zonoid depth in $\text{conv}(\mathbf{X}_{obs}^{(n)})$, $\exists D_{\alpha(n)}(\mathbf{X}_{obs}^{(n)})$ such that $\mathbf{y} \in \partial D_{\alpha(n)}(\mathbf{X}_{obs}^{(n)})$. As $n \rightarrow \infty$ $\partial D_{\alpha(n)}(\mathbf{X}_{obs}^{(n)}) \xrightarrow{a.s.} \partial D_{\alpha}(\mathcal{E})$ for some α such that $\mathbf{y} \in \partial D_{\alpha}(\mathcal{E})$, which is an ellipsoid described by $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and a scaling constant.

For Tukey depth, regard the sequence of the regions $D_{\alpha(n)}(\mathbf{X}_{obs}^{(n)})$ with $\alpha(n)$ defined by (4). Due to absolute continuity and ellipticity of \mathcal{E} , as $n \rightarrow \infty$ $\partial D_{\alpha(n)}(\mathbf{X}_{obs}^{(n)}) \xrightarrow{a.s.} \partial D_{\alpha}(\mathcal{E})$, an ellipsoid described by $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, and containing \mathbf{y} . \square

Proof of Theorem 2:

The proof for Mahalanobis depth is obvious.

For zonoid and Tukey depth, due to consistency with the population version, we restrict to the population one. Let $Y \sim \mathcal{E}$ and allow for a transform $Y \mapsto Z = \mathbf{R}\Sigma^{-1/2}(X - \mu)$ with \mathbf{R} being a rotation operator such that w.l.o.g. $\mathbf{y} \mapsto \mathbf{z}$ and $\mathbf{z}_i = 0 \forall i = 2, \dots, d$ with d being the missing values orthant. We have to show that $D(\mathbf{z}|Z) > D(\mathbf{x}|Z) \forall \mathbf{x}$ with $\mathbf{x}_i = \mathbf{z}_i \forall i = 1, \dots, d-1$. Let $D(\mathbf{z}|Z) = \alpha$, and regard the corresponding region $D_\alpha(Z)$. As both zonoid and Tukey depths satisfy the weak projection property, due to Statement 1 of Theorem 2 from [Dyckerhoff \(2004\)](#) it is sufficient to check univariate projections in the plane spanned by orthants \mathbf{e}_1 and \mathbf{e}_d , namely only angles in range $(0, \pi/2]$ between \mathbf{e}_1 and the direction $\mathbf{u} \in \mathbb{S}^2$, taken counter clockwise, say. $F_{\mathbf{u}|Z}(x) = (1 - p_{NA})F_1(x) + p_{NA}F_1(x/\cos\beta)$ for $\beta \in [0, \pi/2]$ with $\beta = \arccos(\mathbf{u}'\mathbf{e}_1)$, $F_1(x)$ being a univariate marginal c.d.f. of Z assuming it has no missing values, and p_{NA} is the portion of missing values. $F_{\mathbf{u}|Z}(x) > F_1(x) = F_{\mathbf{e}_1|Z}(x) \forall \beta \in (0, \pi/2]$ for $x > 0$ ('<' for $x < 0$), and this last equality holds for ray pointing at the imputed point. Due to monotonicity, the inequalities reverse for the quantile function, which combined with the definitions of the both depths gives strictly $D_\alpha(\mathbf{u}|Z) \subset D_\alpha(\mathbf{e}_1|Z) \forall \alpha \in (0, 1)$ and $\forall \beta \in (0, \pi/2]$. \square

Proof of Proposition 1: Factorization of the determinant and of the inverse gives $|\mathbf{A} + \mathbf{a}\mathbf{a}'| = |\mathbf{A}|(1 + \mathbf{a}'\mathbf{A}^{-1}\mathbf{a})$, $(\mathbf{A} + \mathbf{a}\mathbf{a}')^{-1} = \mathbf{A}^{-1} - \frac{(\mathbf{A}^{-1}\mathbf{a})(\mathbf{a}'\mathbf{A}^{-1})}{1 + \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}}$, see Appendix A by [Mardia et al. \(1979\)](#), further used in proofs of Lemma 1 and of Proposition 2):

$$\begin{aligned} |\mathbf{Y}'\mathbf{Y}| &= |\mathbf{X}'\mathbf{X} + \mathbf{y}_k\mathbf{y}_k' - \mathbf{x}_k\mathbf{x}_k'| = |\mathbf{X}'\mathbf{X} + \mathbf{y}_k\mathbf{y}_k'| (1 - \mathbf{x}_k'(\mathbf{X}'\mathbf{X} + \mathbf{y}_k\mathbf{y}_k')^{-1}\mathbf{x}_k) \\ &= |\mathbf{X}'\mathbf{X}| (1 + \mathbf{y}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{y}_k) \left(1 - \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k + \frac{\mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{y}_k\mathbf{y}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k}{1 + \mathbf{y}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{y}_k} \right) \\ &= |\mathbf{X}'\mathbf{X}| (1 + \mathbf{y}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{y}_k - \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k - \mathbf{y}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{y}_k\mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k \\ &\quad + \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{y}_k\mathbf{y}_k'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_k). \end{aligned}$$

Let $\mathbf{z}_k = \mathbf{x}_k - \mathbf{y}_k$, then due to Mahalanobis orthogonality of \mathbf{y}_k and \mathbf{z}_k w.r.t. \mathbf{X} ($\mathbf{y}_k'\Sigma_{\mathbf{X}}^{-1}\mathbf{z}_k = 0$), one has

$$d_M^2(\mathbf{x}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) = d_M^2(\mathbf{y}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) + d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}).$$

Using this, one obtains

$$|\mathbf{Y}'\mathbf{Y}| = |n\Sigma_{\mathbf{X}}| \left(1 - \frac{1}{n}d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \left(1 + \frac{1}{n}d_M^2(\mathbf{y}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \right) \right).$$

As $d_M^2(\mathbf{x}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) < n$ one has

$$\frac{1}{n}d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \left(1 + \frac{1}{n}d_M^2(\mathbf{y}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \right) < \frac{1}{n}d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \left(2 - \frac{1}{n}d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \right).$$

Further, one can rewrite the right term as

$$\frac{1}{n}d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \left(2 - \frac{1}{n}d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) \right) = \frac{1}{n^2}g(d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}})),$$

with $g(x) = -x^2 + 2nx$, a function monotonically increasing from $g(0) = 0$ to $g(n) = n^2$, while $d_M^2(\mathbf{z}_k, \mu_{\mathbf{X}}; \Sigma_{\mathbf{X}}) > 0$ as long as $\mathbf{y}_k \neq \mathbf{x}_k$. From this follows that $|\mathbf{Y}'\mathbf{Y}| < |n\Sigma_{\mathbf{X}}|$, and thus $|\Sigma_{\mathbf{Y}}| < |\Sigma_{\mathbf{X}}|$. \square

Proof of Lemma 1: W.l.o.g. we restrict to the case $i = 1$. Let \mathbf{Z} be a $n \times d$ matrix with $\boldsymbol{\mu}_{\mathbf{Z}} = \mathbf{0}$ and $\mathbf{z}_{1,miss(1)} = \boldsymbol{\Sigma}_{\mathbf{Z} \text{ miss}(1), \text{obs}(1)} \boldsymbol{\Sigma}_{\mathbf{Z} \text{ obs}(1), \text{obs}(1)}^{-1} \mathbf{z}_{1, \text{obs}(1)}$. Denoting $\mathbf{a} = (0, \dots, 0, \mathbf{y}')' \in \mathbb{R}^d$, then

$$\boldsymbol{\Sigma}_{\mathbf{Z}} = \mathbf{Z}'\mathbf{Z} - \mathbf{z}_1 \mathbf{z}_1' + (\mathbf{z}_1 + \mathbf{a})(\mathbf{z}_1 + \mathbf{a})' - \frac{1}{n} \mathbf{a} \mathbf{a}'.$$

As $\mathbf{z}_1 \boldsymbol{\Sigma}_{\mathbf{Z}}^{-1} \mathbf{a} = 0$ due to Mahalanobis orthogonality, by simple algebra (analog to that in the proof of Lemma 1)

$$|\boldsymbol{\Sigma}_{\mathbf{Z}}| = |\boldsymbol{\Sigma}_{\mathbf{Z}}| \left(1 + \frac{n-1}{n^2} \mathbf{a} \boldsymbol{\Sigma}_{\mathbf{Z}}^{-1} \mathbf{a}\right).$$

□

Proof of Proposition 2: Factorization of the determinant and of the inverse gives:

$$|\mathbf{Y}'\mathbf{Y}| = |\mathbf{X}'\mathbf{X}| \left((1 - \mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k)^2 + \mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{z}_k (1 - \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k) \right)$$

with $\mathbf{z}_k = \mathbf{x}_k - \mathbf{y}_k$. Let \mathbf{M} be a $(d \times d)$ matrix with elements $M_{i,i} = 1$ for $i \in \text{miss}(k)$ and zero otherwise. Then $\mathbf{z}_k = \sigma^2 \mathbf{M} \mathbf{V} \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{u}_k$, and thus, denoting $\mathbf{u}_k^* = \sigma \mathbf{M} \mathbf{V} \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{u}_k$,

$$\mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{z}_k = (\mathbf{u}_k^*)' \mathbf{V} \boldsymbol{\Lambda}^{-\frac{1}{2}} \sigma^2 \boldsymbol{\Lambda}^{-\frac{1}{2}} \mathbf{V}' \mathbf{u}_k^* < (\mathbf{u}_k^*)' \mathbf{u}_k^* = \mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k,$$

because $\sigma^2 < \lambda_d$.

In the same way one can show that $\mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k < \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k$. Keeping

$$0 < \mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{z}_k < \mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k < \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k < 1,$$

clearly $|\mathbf{Y}'\mathbf{Y}| < |\mathbf{X}'\mathbf{X}| ((1-a)^2 + a(1-x))$ with $a = \mathbf{z}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k$ and $x = \mathbf{x}_k'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_k$, and this last term is a function of a and x , that is < 1 for all $(a, x) \in (0, 1)^2$. □

Proof of Theorem 3: First point can be checked by elementary algebra. Second point follows from the coordinatewise application of Lemma 1. For third point it suffices to prove the single-output regression case. The regularized PCA algorithm will converge if

$$\mathbf{y}_{id} = \sum_{s=1}^d u_{is} \sqrt{\lambda_s} v_{ds} = \sum_{s=1}^d u_{is} \left(\sqrt{\lambda_s} - \frac{\sigma^2}{\sqrt{\lambda_s}} \right) v_{ds}$$

for any $\sigma^2 \leq \lambda_d$. W.l.o.g. we prove that

$$\mathbf{y}_d = \boldsymbol{\Sigma}_{d(1, \dots, d-1)} \boldsymbol{\Sigma}_{(1, \dots, d-1)(1, \dots, d-1)}^{-1} \mathbf{y}_{(1, \dots, d-1)} \iff \sum_{i=1}^d \frac{u_i v_{di}}{\sqrt{\lambda_i}} = 0$$

denoting $\boldsymbol{\Sigma}(\mathbf{Y})$ simply $\boldsymbol{\Sigma}$ for the centered \mathbf{Y} and an arbitrary point \mathbf{y} . Due to the matrix algebra

$$\begin{aligned} \mathbf{y}_d &= \boldsymbol{\Sigma}_{d(1, \dots, d-1)} \boldsymbol{\Sigma}_{(1, \dots, d-1)(1, \dots, d-1)}^{-1} \mathbf{y}_{(1, \dots, d-1)} \\ &= -((\boldsymbol{\Sigma}^{-1})_{dd})^{-1} (\boldsymbol{\Sigma}^{-1})_{d(1, \dots, d-1)} \mathbf{y}_{(1, \dots, d-1)}, \\ \sum_{i=1}^d u_i \sqrt{\lambda_i} v_{di} &= - \left(\sum_{i=1}^d \frac{v_{di}^2}{\lambda_i} \right)^{-1} \left(\sum_{i=1}^d \frac{v_{di} v_{1i}}{\lambda_i}, \sum_{i=1}^d \frac{v_{di} v_{2i}}{\lambda_i}, \dots, \sum_{i=1}^d \frac{v_{di} v_{(d-1)i}}{\lambda_i} \right) \times \\ &\quad \times \left(\sum_{i=1}^d u_i \sqrt{\lambda_i} v_{1i}, \sum_{i=1}^d u_i \sqrt{\lambda_i} v_{2i}, \dots, \sum_{i=1}^d u_i \sqrt{\lambda_i} v_{(d-1)i} \right)'. \end{aligned}$$

After reordering terms one obtains

$$\sum_{i=1}^d u_i \sqrt{\lambda_i} \sum_{j=1}^d \frac{v_{dj}}{\lambda_j} \sum_{k=1}^d v_{ki} v_{kj} = 0.$$

Due to the orthogonality of \mathbf{V} , $d^2 - d$ terms from the two outer sum signs are zero. Gathering nonzero terms, *i.e.* those with $i = j$ only

$$\sum_{i=1}^d u_i \sqrt{\lambda_i} \frac{v_{di}}{\lambda_i} = \sum_{i=1}^d \frac{u_i v_{di}}{\sqrt{\lambda_i}} = 0.$$

□

Derivation of (7): The integrated quantity is the conditional depth density that can be obtained from the joint one by the volume transformation (denoting $d_{Mah.}(z, \boldsymbol{\mu})$ Mahalanobis distance between a point of depth z and $\boldsymbol{\mu}$):

$$\begin{aligned} f_{D((X|X_{obs}=\mathbf{x}_{obs})|X)}(z) &= f_{D(X|X)}(z) \cdot C \cdot T_{down}(d_{Mah.}(z, \boldsymbol{\mu})) \cdot T_{up}(d_{Mah.}(z, \boldsymbol{\mu}^*)) \times \\ &\quad \times T_{angle}(d_{Mah.}(z, \boldsymbol{\mu}), d_{Mah.}(z, \boldsymbol{\mu}^*)). \end{aligned}$$

Any constant C is neglected, as it is unimportant when drawing. The three terms correspond to descaling density to dimension one (downscaling), re-scaling it to the dimension of missing values (upscaling), and linear change of its volume to the hyperplane of missingness (angle transformation).

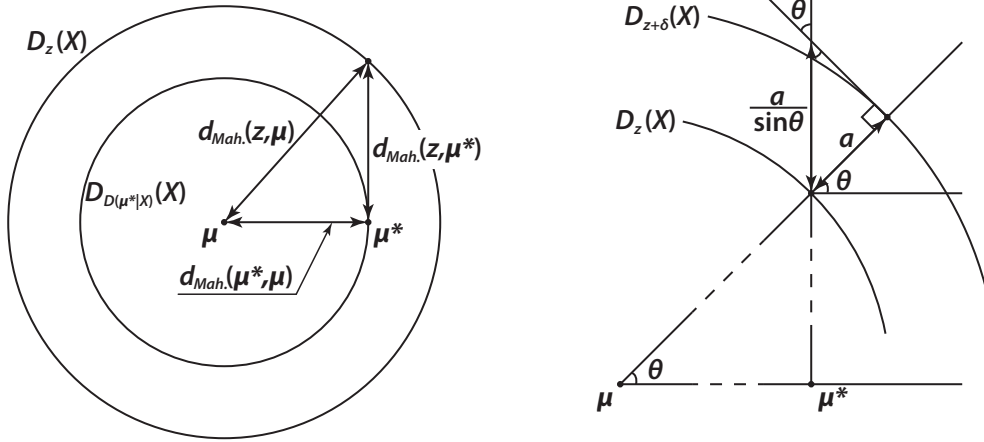


Figure 8: Illustration of the derivation of (7).

$$\begin{aligned}
T_{down}(d_{Mah.}(z, \boldsymbol{\mu})) &= d_{Mah.}^{1-d}(z, \boldsymbol{\mu}) = \frac{1}{d_{Mah.}^{d-1}(z, \boldsymbol{\mu})} . \\
T_{up}(d_{Mah.}(z, \boldsymbol{\mu}^*)) &= d_{Mah.}^{|miss(\mathbf{x})|-1}(z, \boldsymbol{\mu}^*) \\
&= \left(\sqrt{d_{Mah.}^2(z, \boldsymbol{\mu}) - d_{Mah.}^2(D(\boldsymbol{\mu}^*|X), \boldsymbol{\mu})} \right)^{|miss(\mathbf{x})|-1} . \\
T_{angle}(d_{Mah.}(z, \boldsymbol{\mu}), d_{Mah.}(z, \boldsymbol{\mu}^*)) &= \frac{1}{\sin \theta} = \frac{1}{\frac{d_{Mah.}(z, \boldsymbol{\mu}^*)}{d_{Mah.}(z, \boldsymbol{\mu})}} \\
&= \frac{d_{Mah.}(z, \boldsymbol{\mu})}{\sqrt{d_{Mah.}^2(z, \boldsymbol{\mu}) - d_{Mah.}^2(D(\boldsymbol{\mu}^*|X), \boldsymbol{\mu})}} .
\end{aligned}$$

T_{down} and T_{up} are illustrated in Figure 8 (left), for T_{angle} see Figure 8 (right). Letting $d_{Mah.}(z, \boldsymbol{\mu}) = d_{Mah.}(z)$ to shorten notation gives (7).

References

- Bazovkin, P. and Mosler, K. (2015), ‘A general solution for robust linear programs with distortion risk constraints’, *Annals of Operations Research* **229**(1), 103–120.
- Bertsekas, P. D. (1999), *Nonlinear programming. Second edition*, MIT Press.
- Cascos, I. and Molchanov, I. (2007), ‘Multivariate risks and depth-trimmed regions’, *Finance and Stochastics* **11**(3), 373–397.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), ‘Maximum likelihood from incomplete data via the em algorithm’, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), 1–38.
- Donoho, D. L. and Gasko, M. (1992), ‘Breakdown properties of location estimates based on halfspace depth and projected outlyingness’, *The Annals of Statistics* **20**(4), 1803–1827.
- Dyckerhoff, R. (2004), ‘Data depths satisfying the projection property’, *Advances in Statistical Analysis* **88**(2), 163–190.
- Dyckerhoff, R. and Mosler, K. (2011), ‘Weighted-mean trimming of multivariate data’, *Journal of Multivariate Analysis* **102**, 405–421.
- Dyckerhoff, R. and Mozharovskiy, P. (2016), ‘Exact computation of the halfspace depth’, *Computational Statistics and Data Analysis* **98**, 19–30.
- Efron, B. (1994), ‘Missing data, imputation, and the bootstrap’, *Journal of the American Statistical Association* **89**(426), 463–475.
- Efron, B. and Morris, C. (1972), ‘Empirical bayes on vector observations: An extension of stein’s method’, *Biometrika* **59**(2), 335–347.
- Fang, K., Kotz, S. and Ng, K. (1990), *Symmetric multivariate and related distributions*, Monographs on statistics and applied probability, Chapman and Hall.

- Hallin, M., Paindaveine, D. and Šiman, M. (2010), ‘Multivariate quantiles and multiple-output regression quantiles: From l_1 optimization to halfspace depth’, *The Annals of Statistics* **38**(2), 635–669.
- Hastie, T., Mazumder, R., Lee, D. J. and Zadeh, R. (2015), ‘Matrix completion and low-rank svd via fast alternating least squares’, *Journal of Machine Learning Research* **16**, 3367–3402.
- Healy, M. and Westmacott, M. (1956), ‘Missing values in experiments analysed on automatic computers’, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **5**(3), 203–206.
- Jörnsten, R. (2004), ‘Clustering and classification based on the $\{L1\}$ data depth’, *Journal of Multivariate Analysis* **90**(1), 67–89. Special Issue on Multivariate Methods in Genomic Data Analysis.
- Josse, J. and Husson, F. (2012), ‘Handling missing values in exploratory multivariate data analysis methods’, *Journal de la Société Française de Statistique* **153**(2), 79–99.
- Koshevoy, G. and Mosler, K. (1997), ‘Zonoid trimming for multivariate distributions’, *The Annals of Statistics* **25**(5), 1998–2017.
- Lange, T., Mosler, K. and Mozharovskiy, P. (2014), ‘Fast nonparametric classification based on data depth’, *Statistical Papers* **55**(1), 49–69.
- Little, R. and Rubin, D. (2002), *Statistical analysis with missing data*, Wiley series in probability and mathematical statistics. Probability and mathematical statistics, Wiley.
- Liu, R. Y., Parelius, J. M. and Singh, K. (1999), ‘Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by liu and singh)’, *The Annals of Statistics* **27**(3), 783–858.
- Liu, R. Y. and Singh, K. (1993), ‘A quality index based on data depth and multivariate rank tests’, *Journal of the American Statistical Association* **88**(401), 252–260.
- Mahalanobis, P. C. (1936), ‘On the generalised distance in statistics’, **2**(1), 49–55.
- Mardia, K., Kent, J. and Bibby, J. (1979), *Multivariate analysis*, Probability and mathematical statistics, Academic Press.
- Mosler, K. (2002), *Multivariate Dispersion, Central Regions, and Depth: The Lift Zonoid Approach*, Lecture Notes in Statistics, Springer New York.
- Mosler, K. (2013), Depth statistics, in C. Becker, R. Fried and S. Kuhnt, eds, ‘Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather’, Springer. Berlin, pp. 17–34.
- Paindaveine, D. and Bever, G. V. (2013), ‘From depth to local depth: A focus on centrality’, *Journal of the American Statistical Association* **108**(503), 1105–1119.
- Reiter, J. P. and Raghunathan, T. E. (2007), ‘The multiple adaptations of multiple imputation’, *Journal of the American Statistical Association* **102**(480), 1462–1471.

- Rousseeuw, P. J. and Van Driessen, K. (1999), 'A fast algorithm for the minimum covariance determinant estimator', *Technometrics* **41**(3), 212–223.
- Rubin, D. B. (1996), 'Multiple imputation after 18+ years', *Journal of the American Statistical Association* **91**(434), 473–489.
- Schafer, J. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press.
- Seaman, S., Galati, J., Jackson, D. and Carlin, J. (2013), 'What is meant by missing at random?', *Statistical Science* **28**(2), 257–268.
- Stekhoven, D. J. and Bühlmann, P. (2012), 'MissForest – non-parametric missing value imputation for mixed-type data.', *Bioinformatics* **28**(1), 112–118.
- Templ, M., Kowarik, A. and Filzmoser, P. (2011), 'Iterative stepwise regression imputation using standard and robust methods', *Computational Statistics and Data Analysis* **55**(10), 2793–2806.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001), 'Missing value estimation methods for dna microarrays', *Bioinformatics* **17**(6), 520–525.
- Tukey, J. W. (1974), Mathematics and the Picturing of Data, in R. D. James, ed., 'International Congress of Mathematicians 1974', Vol. 2, pp. 523–532.
- van Buuren, S. (2012), *Flexible Imputation of Missing Data (Chapman & Hall/CRC Interdisciplinary Statistics)*, Chapman and Hall/CRC.
- Vardi, Y. and Zhang, C.-H. (2000), 'The multivariate L1-median and associated data depth', *Proceedings of the National Academy of Sciences* **97**(4), 1423–1426.
- Yeh, I.-C., Yang, K.-J. and Ting, T.-M. (2009), 'Knowledge discovery on RFM model using bernoulli sequence', *Expert Systems with Applications* **36**(3, Part 2), 5866–5871.
- Zuo, Y. and Serfling, R. (2000), 'General notions of statistical depth function', *The Annals of Statistics* **28**(2), 461–482.